

Mapping the Landscape of UK Health Data Research & Innovation

A snapshot of activity in 2017

Dr Ekaterini Blaveri

October 2017

CONTENTS

Foreword	5
1. Background to this review	7
1.1. Aims	8
1.2. Scope and sections	8
1.3. Approach	9
1.4. Challenges	10
1.5. Acknowledgements	10
2. Overview of major informatics investments by Funder	11
2.1. MRC	12
2.1.1. The Farr Institute of Health Informatics Research	12
2.1.2. MRC Medical Bioinformatics initiative	16
2.1.3. MRC Capital Investment in Genomics England Data Infrastructure	19
2.2.1. The NIHR Biomedical Research Centres	21
2.2.2. The NIHR Health Informatics Collaborative programme	22
2.2.3. The NIHR Collaboration for Leadership in Applied Health Research and Care	25
2.3. Department of Health and Social Care	26
2.3.1. Academic Health Science Centres	26
2.3.2. Genomics England 100,000 Genomes Project	27
2.4. NHS	28
2.4.1. Academic Health Science Networks	28
2.4.2. NHS Genomic Medicine Centres	30
2.4.3. NHS Global Digital Exemplars	31
2.5. ESRC's investments in Big Data	32
2.5.1. Administrative Data Research Network	32
2.6. EPSRC's health data research investments	33
3. Key informatics investments by Research Organisation	35
England – East Anglia and Midlands	38
University of Cambridge	41
EMBL-European Bioinformatics Institute	48
Wellcome Sanger Institute	52
University of Birmingham	56
University of Leicester	62
University of Warwick	68
England – London	74
Imperial College London	76
King's College London	85
London School of Hygiene and Tropical Medicine	94
Queen Mary University of London	98
University College London	103

England – North	110
University of Leeds	112
University of Liverpool	118
University of Manchester	125
Newcastle University	133
University of Sheffield	141
England – South	147
University of Oxford	148
University of Bristol	156
University of Southampton	163
University of Surrey	170
Northern Ireland	176
Queen’s University Belfast	177
Scotland	182
University of Dundee	185
University of Edinburgh	191
University of Glasgow	199
Wales	204
Cardiff University	207
Swansea University	213
ANNEXES	221
Annex 1	222
Biomedical Research Centres research themes	222
Annex 2	226
Further information on ESRC investments in Big Data	226
Annex 3	228
List of abbreviations	228

FOREWORD

Over the past decade, the Medical Research Council (MRC) and other major research funders have recognised the huge potential of health data research. Through careful and considerable investment, the funders have collectively built capacity in the sector; and enabled a dramatic change in the use of large patient data and research datasets to help understand, treat and prevent disease. The MRC has prioritised strategic support to this domain of activity, with key investments including: the Farr Institute of Health Informatics Research; MRC Medical Bioinformatics awards: contribution to Genomics England Ltd; and most recently, leading the establishment of the new national institute for health data science, Health Data Research UK (HDR UK).

Together with funding from government, universities and charities, these investments have provided a significant step-change in the UK health data landscape across the spectrum, from molecules to man. However, many of these investments are at risk of operating in isolation from each other, with data, knowledge and methods tending to be developed in silos, particularly between the molecular and health informatic fields. Furthermore, the sector still has significant methodological challenges and digital skills shortages. As the volume and complexity of data continues to increase, it's clear that we need to build on UK expertise, enhance our cutting edge analytical methods, and take a more coordinated approach to managing our unique data resources.

Until now, there has been no UK overview of the health data research landscape. The report *Mapping the Landscape of UK Health Data Research and Innovation* gives a unique window into the major investments made across 26 research organisations throughout the UK.

This report was prepared by the MRC to support the funders' investment decisions into HDR UK. However, it also underscores the importance of health data research for the nation in general – and for science in particular – and will help funders develop a more coordinated and mutually beneficial approach, with an aligned set of delivery mechanisms for future investment.

This is a transformational time for the health data research community, with significant activity across the world. The UK not only has a tremendous wealth and variety of health data but also a real strength in analytical skills. This report reveals that the UK has a very real opportunity to be a global leader in health data research and by harnessing this potential could yield huge benefits for our nation's health and wealth.

A handwritten signature in black ink that reads "Rob Buckle". The signature is written in a cursive style with a long, sweeping underline that extends to the right.

Dr R S Buckle
Chief Science Officer, MRC



CHAPTER 1 BACKGROUND TO THIS REVIEW

1.1. Aims

As part of the activities to establish Health Data Research UK (HDR UK) – previously known as the UK Health and Biomedical Informatics Research Institute – the MRC conducted a health data research and innovation landscape mapping review with the following aims:

- To create a broad baseline of current activity in health data research relevant to health/biomedicine and innovation programmes;
- To inform the next phase of investments in this area and in particular the new informatics institute (HDR UK), and other related activities, such as opportunities arising from the [Life Sciences Industrial Strategy](#);
- To compile qualitative insights that will help create a common understanding of the purpose, likely direction, and impacts of existing investments in health data research across different stakeholders that will result in strengthened partnerships and closer alignment of common objectives across existing and new initiatives; and
- To present to the HDR UK Director, and the independent Expert Review Group (ERG), a body of knowledge to help them develop a comprehensive research strategy for HDR UK, including the selection of the research organisation sites that will constitute the institute.

1.2. Scope and sections

Scope

This review focuses on health data research and innovation investments and activities in the UK, and mainly in England due to the complexity and fragmentation across the breadth of activities in England. It was agreed with the sponsors of this review that if required, it could be broadened to include further investments (e.g. international), and/or deepened to include further analysis on specific investments at a future stage.

Sections

In its first section, the review provides a top-level view of some **key informatics/health data science-related** investments from several funders including the:

- **MRC:** Farr Institute of Health Informatics Research; and Medical Bioinformatics awards
- **NIHR:** Health Informatics Collaborative (HIC); and other informatics activities supported through the NIHR Biomedical Research Centres
- **Department of Health and Social Care England:** Academic Health Science Centres
- **NHS England:** Academic Health Science Network; and the National 100,000 Genomes Project
- **ESRC:** Administrative Data Research Centres and Network; and other investments in Big Data
- **EPSRC:** Alan Turing Institute and Health Data Analytics Network

In its second section, the review provides an insight analysis of the **confidential Statements of Interest (Sol)** submitted from the different Research Organisations (ROs) in response to a targeted call issued in December 2016 to express their interest to become an HDR UK substantive or partner site.

Out of scope

This review does not include activities or investments on:

- computational tools and platforms developed by the digital healthcare industry;
- detailed description of data resources and population cohorts; and
- information governance and data access processes and policies.

1.3. Approach

The review staff (Dr Ekaterini Blaveri) used a combination of desk research and stakeholder communications to gain an understanding of the key UK programmes in health data research. Going forward the use of surveys, structured interviews; and workshop(s) with key opinion leaders could be considered to gain a deeper understanding of certain activities.

It should be noted that the material searched could potentially be much more extensive, since aspects of research collaboration are referenced over a wide range of publications. However, because of the limited time available to this review, information was collated predominately from the funders of key programmes.

Extracting insights from the Statements of Interest received from UK Research Organisations in response to the December 2016 call

The MRC office received Sols from 26 ROs interested in becoming a substantive site of HDR UK. These were subdivided into seven geographical clusters, to make easier to identify the regional strengths and interconnections.

As the information requested from the ROs was very high-level, a modified “SWOT” analysis – (Strengths, Translational and Innovation Capabilities, Opportunities and Gaps – STOG) was performed to synthesise meaningful insights from the Sol and address the following questions:

Strengths

- What the organisation already does well?
- What resources (both tangible and intangible) do they have that will help them become a substantive site?
- Have they developed a leading position in health data science-relevant areas – based on what?

Translation and Innovation capabilities

- What are the resources in place to aid research translation and innovation?
- Do they have strong industry collaborations?
- Do they have links with the NHS?
- Are they innovating in ways that others aren't?

Opportunities

- Are there areas that they have the capability to build on or expand?
- Are there any clear opportunities for collaboration and synergies?

Potential gaps

- Which informatics/health data science areas they have not developed?
- Is there a perception that hampers its ability to succeed?

Any information that was deemed secondary to the “SOTIG” analysis, but was still a valuable piece of evidence, was broadly grouped into:

Research themes: research domains (e.g. discovery science, digital health, stratified medicine and population and public health) and disease areas;

Analytics: methodology innovation and new tools;

e-Infrastructure: platforms/environments/systems and compute to enable data usage and support collaboration; and

People: public engagement, training and partnerships/collaborations.

Where possible, the information provided within the Sol was cross-checked with other sources.

1.4. Challenges

Scoping the UK (and international) landscape and identifying the different health and biomedical data research and innovation activities and trends is not an easy task.

As this is an emerging field and the discipline is not yet clearly established/defined, identifying what actually qualifies as health and biomedical data research, and what does not, was the first challenge. Another challenge with identifying investments is that not all of them are clearly visible to an external audience. The only thing that can be said with certainty is that no one knows exactly how many investments currently exist in the UK.

Against this background, it is impossible to obtain an exhaustive picture of the current global position regarding health data research investments. This review aims to gain an overview of the main features of the general landscape, rather than to try to fill in every detail. Hence, although this isn't conclusive work, it sets the baseline and will contribute to the on-going discussion and shaping of investment in health data research and innovation, particularly during the early phases of establishing HDR UK. Given the rapid pace of change, there will be an on-going need to refresh and review this landscape – potentially including adapting this report into a web-based community resource.

1.5. Acknowledgements

The information provided in section three of this review was mainly extracted from the Sols of Research Organisations interested in becoming a Substantive Site of the UK Health and Biomedical Informatics Research Institute (now known as HDR UK). This was in response to a targeted call from MRC that preceded the call for the HDR UK Phase 1 investment in Substantive Sites announced in August 2017.

We would like to sincerely thank all the representatives from across the different Research Organisations who have contributed to this report.

Also, we are grateful to the HDR UK funding partners for their input and support in shaping this report.



CHAPTER 2 OVERVIEW OF MAJOR INFORMATICS INVESTMENTS BY FUNDER

2.1. MRC

2.1.1. The Farr Institute of Health Informatics Research

2.1.1.1. The e-Health Informatics Research Centres and Network

In 2012 the MRC brought together 10 government and charity research funders¹ to invest **£19M (£5.8M MRC)** over five years (awards end in April 2018) through four **e-Health Informatics Research Centres (e-HIRCs)** – based at: University College London (UCL), University of Manchester; University of Swansea; and University of Dundee. Collectively the e-HIRCs involve 24 university partners across the UK.

The funders also provided an additional **£2.18M**-award to establish a **UK Health Informatics Research Network (UK HIRN)**. The Network aimed to promote best practice, build partnerships (particularly with academia, industry and the NHS) and co-ordinate activities in outward-facing priority areas including: methodology; information governance; and public engagement.

This joint investment aimed to enhance capability in cutting-edge research in developing informatics tools and methodologies for using and linking e-health data at scale as well as building capacity and skills across a range of areas from cardiovascular to maternal health, using public health, clinical, educational, housing, pollution, accident and demographic datasets.

2.1.1.2. The Farr Institute of Informatics Research

In 2013 the MRC invested a further **£20M** capital (predominately compute infrastructure and buildings) into the four e-HIRCs to establish the [Farr Institute of Health Informatics Research](#) and accelerate the development of a UK-wide platform for health informatics research, including digital infrastructure to securely share data, support linkage at scale and develop standards. After the establishment of the Farr Institute, the four e-HIRCs changed their names to: Farr@London (UCL); Farr@CIPHER (Swansea); Farr@HeRC (Manchester); and Farr@Scotland (Dundee, and later Edinburgh); and the UK HIRN changed its name to be: the Farr Network.

The vision for the Farr Institute was to:

“Create a UK-wide initiative that acts as a catalyst and accelerator for large-scale data science harnessing the knowledge from health records and research data to improve the health of patients and communities across the UK and beyond”.

1. Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates) and the Wellcome Trust

Farr Institute's key activities²

1. Performing pioneering multidisciplinary research with large and complex health-related data

Research excellence is a primary goal of the Farr Institute, ideally through collaborations across the UK so that it acts as a single institute. To this end, it has initiated several **cross-Centre research programmes** (*Box 1*) to facilitate research at scale, and answer questions that the individual Farr Centres or investigators would be unable to pursue in isolation, many with collaborators from other organisations beyond health care research. It's noteworthy that the e-HIRCs and the Farr Institute awards did not include funding to support cross-centre research. Instead, the Farr Institute's Network grant was used to: convene new groups with complementary expertise; and pump-prime collaborative research by reprioritising the budgets of the individual Centres.

Box 1 – Farr Institute's Cross-Centre Projects – Examples

(extracted from the 2016/17 Farr Annual Report):

Health outcomes of children and adults with learning difficulties

This project seeks to investigate factors associated with the diagnosis, treatment, health and educational outcomes of vulnerable groups in society, particularly those with learning difficulties and/or Autistic Spectrum Disorders. A multi-disciplinary team from the Universities of Glasgow, Cardiff and Swansea with expertise in epidemiology, learning disorders and health informatics have been creating comparative cohorts from multiple sources including; education, general practice, inpatient, mortality and population registers. The population cohorts involve more than 1.5 M linked records and the first analysis on relative risk of hospitalization from all and selected causes has commenced.

Identifying episodes of acute kidney injury (AKI) across healthcare settings using routinely collected data

The Kidney Disease@Farr collaboration brings together a multidisciplinary group of researchers with interests in health informatics and kidney disease. This project aimed to evaluate and improve the portability of an existing AKI electronic phenotyping algorithm developed in Grampian, Scotland. The algorithm has now been implemented across three other UK regions: Hampshire, Salford and Swansea. All datasets consisted of linked electronic health record data across primary and secondary care. Using the algorithm, researchers compared the incidence and character of AKI episodes in each region and index year and found substantial differences between regions in AKI incidence. The group are currently systematically investigating potential reasons for these differences. Anticipated outcomes of the project in the next few months are: a publically available AKI phenotyping algorithm, including associated metadata and guidance for local implementation; a clinically focussed paper on incidence and characteristics of AKI episodes in four UK healthcare populations, including potential explanations for differences; a methodological paper providing guidance how to implement electronic phenotyping algorithms across datasets that have different contexts and underlying infrastructures.

2. The information provided in this section was extracted from the Farr Annual Reports. Further information of the Farr institute's activities can be found in other section of this report, under the lead Research Organisations of the individual Farr Centres (University of Edinburgh, University of Manchester, University of Swansea and UCL).

2. Develop new health data research infrastructure, methods and technologies

The Farr Institute has also launched several initiatives to demonstrate secure inter-operability between research data infrastructure investments across the UK, to facilitate data sharing and federated meta-analyses. For example:

- all Farr Centres are actively participating in the [JISC SafeShare project](#) (together with the MRC Medical Bioinformatics initiative and the ESRC Administrative Data Research Network (*see section 2.5.1*)) piloting a high capacity, encrypted network between research centres across the [JANET network](#)³. Swansea University and Cardiff University are the first to complete trialling the new infrastructure. ADRN researchers at these universities are actively using the service for access to sensitive research data.
- MRC's capital investment in the Farr Institute established the **UK Secure eResearch Platform (UKSeRP)**, which is hosted by the Farr@CIPHER(Swansea) and provides ISO27001 certified data curation, management, sharing and analysis facilities to the UK research community. Imbued with capabilities derived from 10 years of operating the Secure Information Linkage (SAIL) Databank in Wales (*see also section in this report describing investments at Swansea University*), UKSeRP is provisioned with a suite of capabilities that allow data owners to remotely share their data in line with best information governance principles using an easy-to-use interface in a ready to use solution. UKSeRP provides underpinning technology for Dementia Platform UK (*Box 2*); the SAIL Databank – its biggest user; UK Biobank; the Administrative Data Research Centre Wales; and the MRC Medical Bioinformatics CLIMB (Cloud Infrastructure for Bioinformatics) programme (*see section 2.1.2*).
- The Farr@HeRC(Manchester) has provided the research and analytical platform for **Connected Health Cities**, a cornerstone of the Northern Powerhouse (*see section in this report describing investments at the University of Manchester*).

Box 2 – The Dementia Platform UK (DPUK) has selected UKSeRP as the unifying Data Portal to bring together and share data from 33 pre-existing research cohort studies (*extracted from the 2016/17 Farr Annual Report*).

With robust legal agreements for data transfer now in place with five academic institutions and an industry partner, data is now being received in Swansea from cohorts across the UK. Three cohorts have currently shared baseline study data to the Portal. Another six are currently online and using our front-end to the infrastructure to upload their data and supporting documentation or are in advanced discussions for sharing multi-omics data. After a short period of testing, this data will be made available to the wider research community, subject to successful project applications. It is envisaged that 16 cohorts which are currently deemed 'research-ready' will have shared data by June 2017 and preliminary feasibility and case study projects are due to start in order to provide operational evidence of the capability and functionality of the DPUK Data Portal. New UKSeRP infrastructure builds include the integration of the DPUK genomics platform to allow for centralised genomic research to take place within a specialist analysis platform on UKSeRP, along with epidemiological, imaging and wearables data. In partnership with the European Medical Information Framework (EMIF), DPUK will be the recipient of extra metadata tools for in-depth cohort discovery, and subsequently, an integrated participant finder tool that will enhance the capability of DPUK cohorts to be used to recruit for clinical trials, which is of particular interest to the European Prevention of Alzheimer's Dementia (EPAD) consortium. UKSeRP has recently added XNAT, an open source imaging informatics platform, to support DPUK's needs, as well as making it available to other tenants.

3. JANET is a high-speed network for the UK research and education community.

3. Work with the owners and data controllers to support the safe use of patient information for medical research across the UK

The Farr Institute recognises that ensuring privacy, security and good governance are paramount. In Wales and Scotland, it has built good relationships with NHS and Government and local authority data controllers. This has been instrumental to facilitate access to data, proportionate governance and improve data linkage. In Scotland, the Farr Institute co-designed the Public Benefits and Privacy Panel launched in 2016. In England, there is senior Farr Institute representation on the NHS Digital Research Advisory Group, and it has made important contributions to the recent [National Data Guardian Review of data security, consent and opt-outs](#).

In terms of the Farr's quest to mobilise new datasets for research, there have been notable developments in national prescribing and national imaging in Scotland and incorporation of local authority and other public datasets in Wales.

4. Training and capacity building

As a research collaboration, the Farr Institute represents a variety of postgraduate (e.g. Masters) and professional education courses (CPD and MOOCs) run and hosted by its member and partner universities and has co-hosted numerous events reaching over 5,000 people. This has created a thriving community of doctoral and post-doctoral researchers across the UK, including over 100 PhD students and the successful launch of the [Farr Institute Future Leaders programme](#), with the aim of training the next generation of leaders in health informatics.

5. Public engagement

The Farr Institute has been very active in public engagement with exposure delivered across science festivals; creation of the [100 Ways case studies series](#); development of the [#datasaveslives](#) campaign; and strategic input to the [Understanding Patient Data](#) initiative led by Wellcome.

6. Partnerships

The Farr Institute further developed existing collaborations and fostered new relationships with diverse groups from across the pharmaceutical and health technology industries (e.g. Intel Life Science, GSK, Jansen) but also government departments and the NHS (e.g. established the [Prudent Healthcare Intelligence Hub](#), an academic/NHS partnership in Wales; and the [Innovative Healthcare Delivery programme](#) in Scotland to realise the vision of harnessing informatics to improve NHS services).

It also continues to establish collaborations across the globe with the leaders of health data science in other countries. For example, it has a memorandum of understanding with the [Institute for Clinical Evaluative Sciences in Canada](#), solidifying a commitment to work together in areas of common strategic operations and interest.

2.1.2. MRC Medical Bioinformatics initiative

In March 2013, the MRC made a £39M (£28M capital for spend in 2014/15; and £11M resource for spend over 5 years of award, ending in 2019) investment to build and extend the UK's capacity, capability and infrastructure in medical bioinformatics by: building flexibly on the capabilities already in place in Universities and Centres of Excellence; developing new infrastructure, tools and increased coordination and sharing of capabilities; and supporting career opportunities for computational scientists, technologists and programme leaders. Fellowship funding was also available for recruitment of research leaders in bioinformatics and computational biology.

This initiative strongly aligned with MRC's existing investments in informatics research, including the EMBL-European Bioinformatics Institute (EBI), [ELIXIR](#), the Farr Institute/e-HIRCs; and those of other funders including those of the Department of Health and Social Care.

The awards were targeted at the “molecule to man” interface and aimed to:

- improve informatics and computational approaches to understanding mechanisms of human disease – from the cellular level through to the whole disease phenotype
- facilitate research linkage between basic genomic information, complex phenotypes, and clinical data sets
- build analytical capabilities tailored to medical research and in particular the challenge of integration across differing degrees of resolution, dynamic scales and heterogeneity of data, and interpretation of very large and rich data sets including genomics, proteomics, metabolomics, imaging, population data.

The following six awards were made:

1. Data-driven discovery for personalised medicine (*short name: eMedLab*)

Crick Institute; University College London (UCL); London School of Hygiene and Tropical Medicine (LSHTM); King's College London (KCL); Queen Mary University of London (QMUL); Sanger Institute; and the EMBL-European Bioinformatics Institute (EMBL-EBI) (*co-leads: Jim Smith and David Lomas. Research lead: Nick Luscombe*) – £8.9M

This MRC investment has established one of the world's largest secure biomedical cloud infrastructures built to integrate and share heterogeneous data from personal healthcare records, imaging, pharmacoinformatics and genomics. Through co-location eMedLab eliminates the delays and security risks that occur when data is moved. It also provides a platform to develop analytical tools that allow biomedical researchers to transform raw data into scientific insights and clinical outcomes. It stores data securely and its modular design ensures sustainability through expansion and replacement.

Currently, eMedLab supports the efforts of over one hundred biomedical researchers from academia and industry, studying cancers, cardio-vascular and rare diseases, including six [MRC Stratified Medicine projects](#) focusing on auto-immune diseases through hosting their tranSMART infrastructure where they have captured and integrated their complex multi-omic data on several thousand rheumatoid arthritis and psoriasis patients.

The system is operated as a private cloud through a Red Hat Enterprise Linux OpenStack Platform with Lenovo Flex System hardware to enable the creation of virtual HPC clusters bespoke to individual researchers' requirements. It is housed in the Virtus DataCentre in Slough, UK through the Jisc Shared Data Centre Agreement.

MRC eMedLab has won an award at the [UK Cloud Awards 2017](#) for the 'Best Public Sector Project' category.

2. Consortium for medical microbial bioinformatics (*short name: CLIMB*)

University of Warwick, Cardiff University, Swansea University, University of Birmingham, University of Bath, University of Leicester and The Quadram Institute (a BBSRC research institute)
(*co-leads: Mark Pallen & Sam Sheppard*) – £8.5M

This investment established what is probably the largest computer (distributed, web-based) system dedicated to the analysis of microbial genomes in the world, the **Cloud Infrastructure for Microbial Bioinformatics** (CLIMB). MRC CLIMB is a public, free to use cloud for the medical microbial bioinformatics community in the UK and currently supports over 1,000 researchers from over 300 research groups across the UK. This state of the art inter-disciplinary facility for use by the academic, industrial and healthcare communities – will enhance regional and national capability and infrastructure in microbial bioinformatics and improve our understanding of bacteria of medical importance.

The system is based on an open source platform for cloud computing powered by OpenStack and provides large scale data storage and compute capacity on demand whilst making these more cost-efficient by optimising the utilisation of hardware while capitalising on economies of scale. CLIMB has allowed UKSeRP to be augmented with significant HPC capabilities to support genetic analyses. Operating fully as part of the UK-wide CLIMB consortium this investment has been accommodated into the existing UKSeRP infrastructure to allow unused cycles to be repurposed for other studies and tenants, as well as providing the necessary secure circumstances to allow the co-analysis of microbial and human genetic data, along with clinical and population data.

MRC CLIMB won two [HPCWire Readers Choice awards](#) in 2017; ‘Best use of HPC in the life sciences’ and ‘Best HPC Collaboration (Academia/Government/Industry)’, which were presented at the international Supercomputing conference in Denver, Colorado.

3. MRC Medical Bioinformatics Centre at the [Leeds Institute for Data Analytics](#)

University of Leeds (*lead: Alex Markham*) – £7.0M

The focus of the Leeds Medical Bioinformatics award is to link molecular, phenotypic and health record datasets, sourced locally and nationally, for research to deliver patient benefit. In this respect this award aligned well with the focus of the Farr Institute.

Their plan is to create an integrated clinical research campus on the Mayo Clinic model where all patients are potential research subjects with all routine phenotype data captured on the Electronic Patient Record (EPR) system and made available in a secure “Phenobank”. The core of this investment is a partnership between the University, the Leeds Teaching Hospitals NHS Trust and the Leeds-based Phoenix Partnership (TTP: the supplier of one of the largest primary care clinical records system in the world).

The award contributed towards upgrading the computing infrastructure (high speed network connections and data storage) that links clinical practice and research activity across the campuses of the University of Leeds and its Teaching Hospitals and establishing a Virtual Research Environment safe haven.

This infrastructure supports researchers studying cancer, inherited rare diseases and acute kidney injury.

4. The Big Data Institute

University of Oxford (*leads: Gil McVean Director, Martin Landray, Deputy Director*) – £6.0M

The Big Data Institute (BDI) lies at the heart of Oxford's ambitions in health data science (*see also section 3 in this report describing investments at BDI@Oxford University*). The MRC Medical Bioinformatics award provided capital funding only that went towards the £47M building, where BDI is housed (opened in May 2017).

BDI has received funding from diverse sources including the UK Government: UK Research Infrastructure Investment Partnership (£10M); the MRC (£1M in addition to the Medical Bioinformatics award); Charity: BHF (£1M); and Philanthropy: Robertson Foundation (£5M capital; £5M fellowships); and Li Ka Shing £20M fellowships).

5. Partnership to develop the aggregation, integration, visualisation and analysis of large, complex data (*short name: UK MEDBIO*)

Imperial College London (*lead: Paul Elliot*) – £5.9M

The focus of this investment was on linking 'omics data to the phenome/exposome⁴ through the work in the MRC-Public Health England (PHE) Centre for Environment and Health. This entailed bringing together the necessary patient/cohort data, new 'omics technologies, computing and other resources, with dedicated infrastructure in data management, integration, analysis and visualisation, to enable major advances in understanding the aetiopathogenesis of chronic human diseases, and the gene-environment-microbiome interactions that underpin them.

This award mainly funded capital infrastructure underpinned by a full-time bioinformatician, system administrator and three five-year fellowships (jointly supported by Imperial College London). It provided extensive computing, data storage and HPC infrastructure that is enabling the analysis of very large datasets (e.g. UK Biobank, GWAS/sequencing, epigenomics, immune data), and for the overlay of multi-omics data (e.g. GWAS/metabolomics). Currently, around 100 projects are using this facility.

6. Improved analysis of population health and biological data

MRC/UVRI Uganda Research Unit (*lead: Pontiano Kaleebu*) – £2.8M

Aligned to the MRC's global health strategy and long term investment in the MRC/UVRI Uganda Research Unit, this MRC Medical Bioinformatics award established the MRC/UVRI Medical Informatics Centre in Entebbe, Uganda. It capitalises on the technological advances in genomics and medical informatics and to ensure the efficient utilisation of data resources in clinical epidemiology and large scale genomics across Sub-Saharan Africa from existing and on-going research programmes.

The Centre builds on established strategic collaborations with UK Research Centres—the Wellcome Trust Sanger Institute; the MRC Functional Genomics Unit; the MRC Centre for Genomics and Global Health; the MRC Epidemiology Unit; School of Biological Sciences, University of Edinburgh; and the MRC Centre for Molecular Bacteriology at Imperial College, UK. These UK Centres provide scientific and logistical support to develop infrastructure and computational processes, cloud computing and human research capacity, as well as supporting pan-African and north-south research partnerships (Africa-UK medical informatics consortium).

4. Refers to the totality of internal and external exposures from a variety of sources including but not limited to chemical and biological agents, gut microbial and psycho-social factors from pre-conception onwards.

Although the MRC Medical Bioinformatics awards were predominately focused on infrastructure, a diverse range of capacity building activities were also supported, including training through MSc and PhD studentships, at least 10 post-doctoral career development fellowships, and several “training academies” covering a breadth of career stages and disciplines, including clinical and non-clinical.

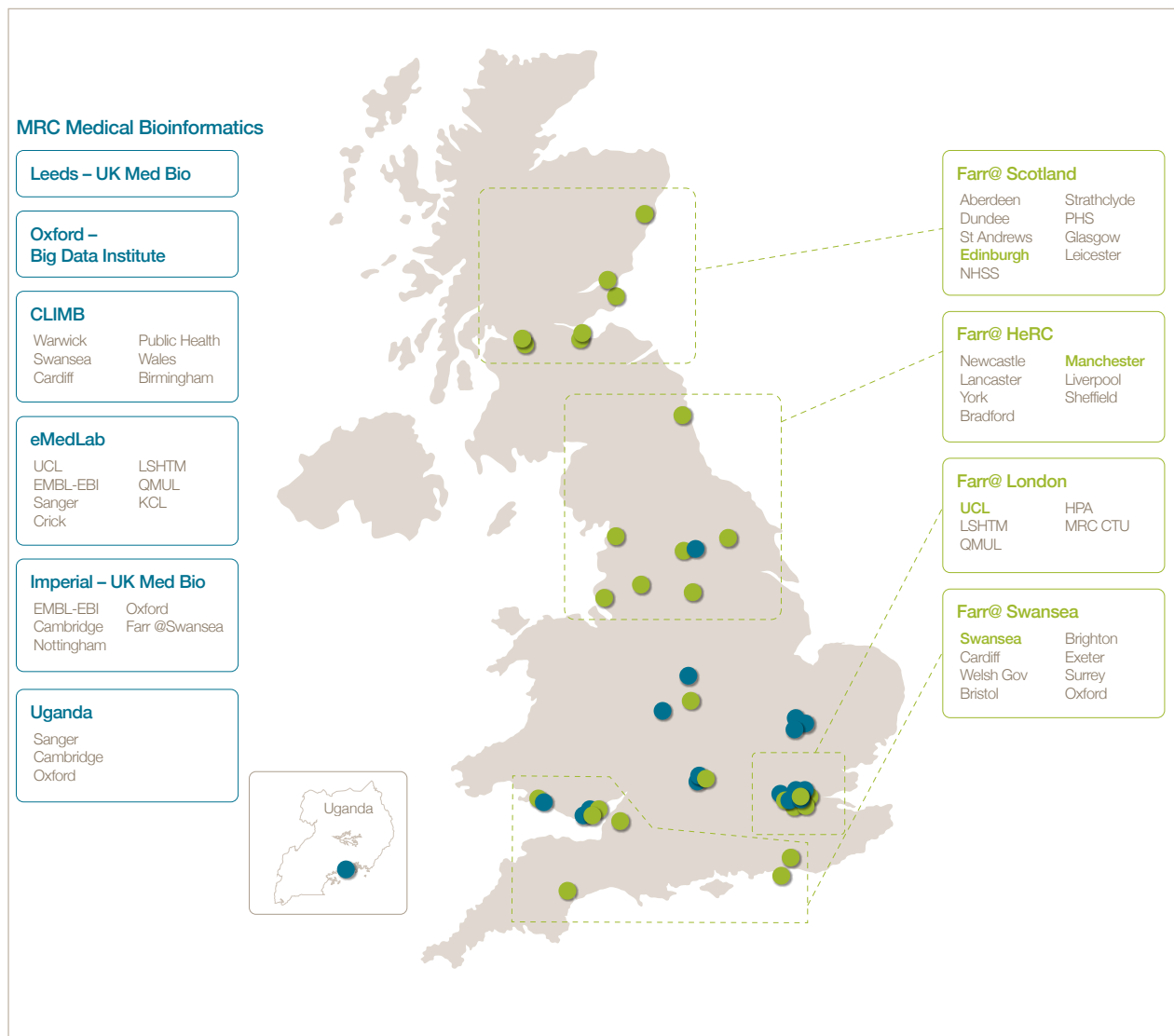
2.1.3. MRC Capital Investment in Genomics England Data Infrastructure

Aligned to the Medical Bioinformatics initiative, in 2014, as part of the MRC-led £230M Clinical Research Infrastructure initiative, £24m capital was invested in Genomics England Ltd (through Queen Mary University London and partners) to support the UK 100,000 Genomes Project. This was used to help establish the main Data Centre and to make the rich data resources securely available to researchers and clinicians from across the UK.

2.1.4. Established infrastructure and capabilities

Over this short period (2012-2014), the MRC, working with its partners invested over £100m (£72M capital, £28M resource) through the initiatives described above. This delivered significant progress towards building the necessary capacity and infrastructure to deliver a step-change in the UK’s health and biomedical data science capabilities. The Research Organisation footprint of these awards is illustrated below.

Figure 1: Farr Institute Centres and MRC Medical Bioinformatics awards



Health Data Research UK, will benefit and build on these capabilities. The individual MRC Medical Bioinformatics and Farr Institute awards will not be renewed, except where they are incorporated into on-going research programmes e.g. the MRC/UVRI Uganda research Unit.

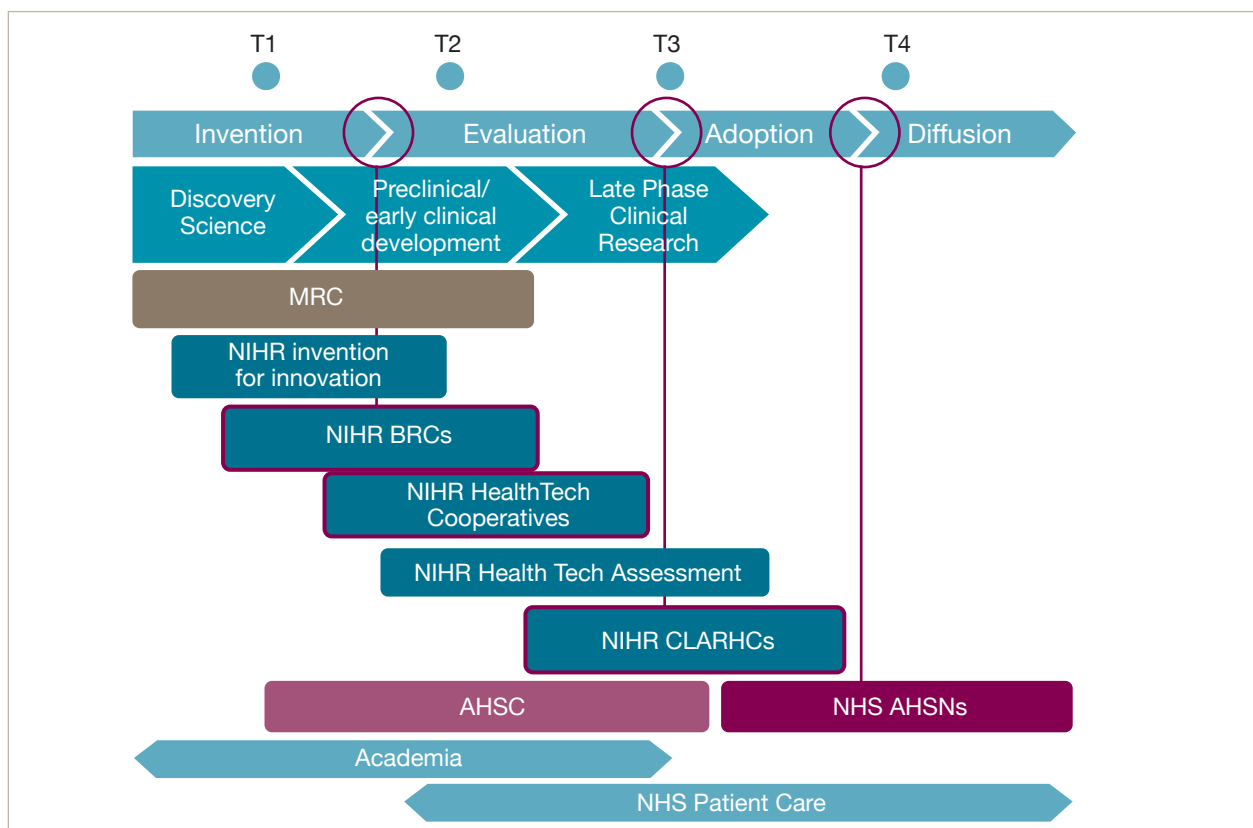
2.2. The National Institute for Health Research

The NIHR is a government body that receives funding from the Department of Health and Social Care (DHSC) in England. Its vision is to improve the health and wealth of the UK through research; and its mission is to provide a health research system in which the NHS supports outstanding individuals working in world-class facilities, conducting leading-edge research focused on the needs of patients and the public.

To do this, the NIHR funds from **early to late stage translational health research** and a range of infrastructure facilities to support it. Furthermore, it works with the patients and the public to shape its research agenda, which spans a number of different areas: from finding new ways of preventing, identifying and treating disease, to evaluating the effectiveness and impact of new treatments, and making sure that the best possible evidence is available to inform decisions about health and social care.

The NIHR and the MRC co-fund several research programmes, which support basic and early translational research. The two bodies work together closely, overseen by the Office for Strategic Coordination of Health Research (OSCHR) to ensure that there is a continuum of research opportunities along the translational pathway. This also applies to informatics/health data research, where there is an **OSCHR health informatics sub-board** (chaired by Professor Sir Alex Markham) to provide an overview of the health informatics landscape including NHS, research and commercial needs, funder alignment and data access. In a time of high activity, the sub-board provides oversight across the community and brings a wide number of organisations operating in this area together.

Figure 2: The transitional pathway for funding health and biomedical informatics and digital health research and innovation programmes



2.2.1. The NIHR Biomedical Research Centres

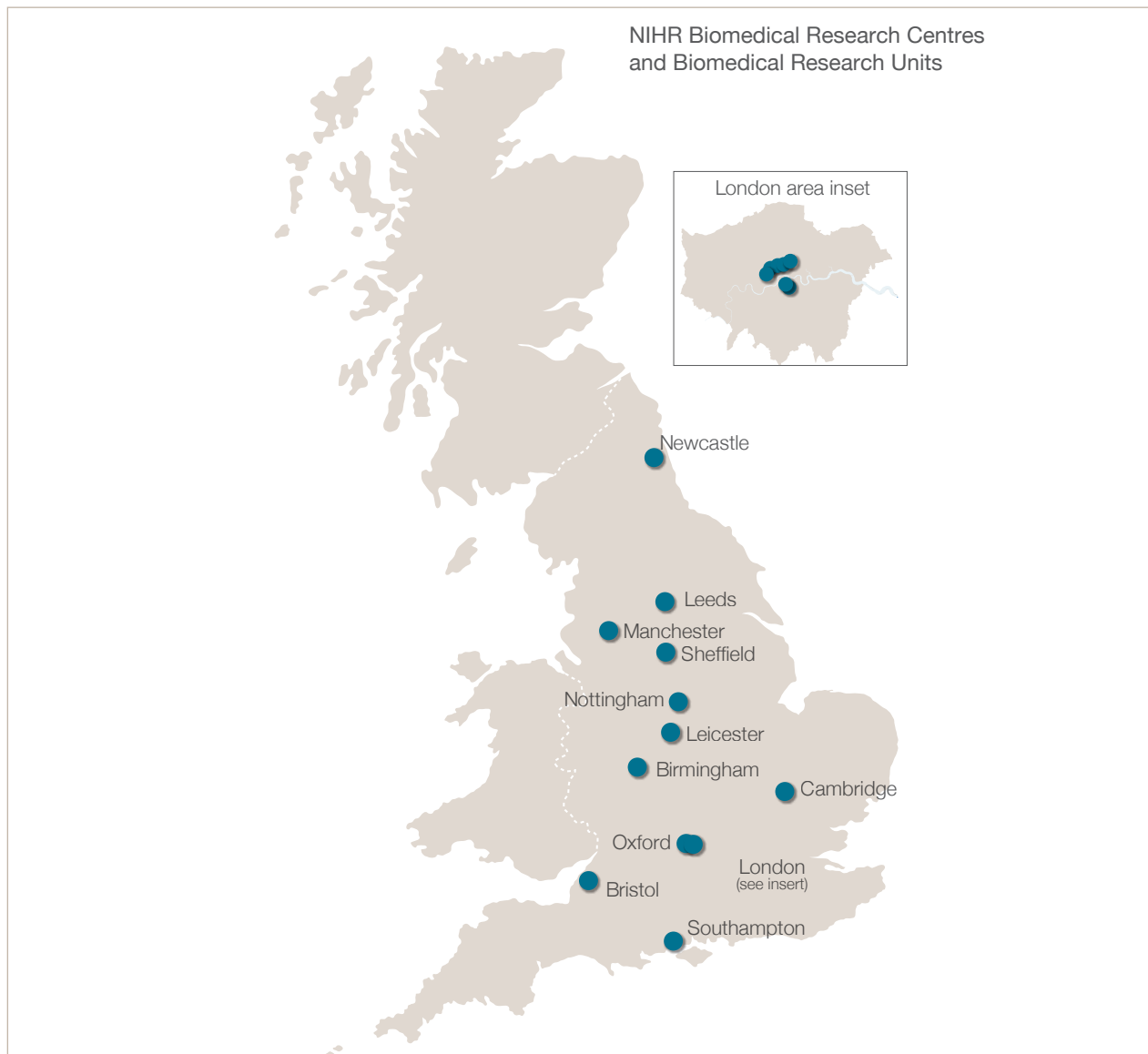
The NIHR infrastructure investments include the Biomedical Research Centres (BRCs), which support several informatics research activities.⁵

The NIHR BRCs are partnerships between: an ‘Academic Partner’, usually a research university or similar specialised institution; and an ‘NHS Host’, usually in the form of an NHS Foundation Trust.

The aims of the BRCs are to:

- drive innovation in the prevention, diagnosis and treatment of ill-health through early translational (experimental medicine) research;
- translate advances in biomedical research into benefits for patients, the health system and for broader economic gain; and
- provide a key component of the NHS contribution to UK’s international competitiveness.

Figure 3: NIHR Biomedical Research Centres distribution across England



5. Please note that details on other NIHR informatics-relevant investments, including the NIHR Bio resource and the NIHR CRIS, are provided in the key informatics investments by Research Organisation (section 3).

The first five BRCs were established in 2007 with the aim to transform scientific breakthroughs into life-saving treatments for patients. Two further five-year cycles of funding followed, with the most recent commencing in April 2017 and allocating £816M across 20 BRCs.

Each BRC has a substantial portfolio of world-class biomedical research across a range of themes and diseases including: cancer, cardiovascular, dementia, endocrinology and metabolism, infection, immunity and inflammation, genetics and genomics, musculoskeletal, neuroscience, nutrition and lifestyle, ageing, paediatrics, respiratory disease, stroke, surgical innovation. For more information see *Annex 1*.

Details on informatics research activities delivered through the BRCs are provided in the sections of this report describing the wider activities of each Research Organisation.

2.2.2. The NIHR Health Informatics Collaborative programme

Background

The NIHR Health Informatics Collaborative (HIC) programme is a stand-alone investment involving more than one BRC. It was conceived in response to a ‘grand challenge’ from England’s Chief Medical Officer: *“to show that data collected in the course of routine patient care across the five largest biomedical research centres can be re-used for collaborative, translational research”*.

The HIC is a collaboration between five leading BRCs. These are:

1. **Oxford BRC:** Oxford University Hospitals NHS Foundation Trust (OUH) and University of Oxford;
2. **Cambridge BRC:** Cambridge University Hospitals NHS Foundation Trust (CUH) and University of Cambridge;
3. **Guy’s & St Thomas’ BRC:** Guy’s & St Thomas’ NHS Foundation Trust (GSTT) and King’s College London;
4. **Imperial BRC:** Imperial College Healthcare NHS Trust (ICHNT) and Imperial College London; and
5. **UCLH BRC:** University College London Hospitals NHS Foundation Trust (UCLH) and University College London.

Three other NHS Trusts are working with the Collaborative:

- Royal Marsden NHS Foundation Trust;
- Kings College Hospital NHS Foundation Trust; and
- Royal Brompton & Harefield NHS Foundation Trust.

Originally, the programme lifecycle was for two years, from April 2013 to March 2015, but this was later extended. The HIC investment was **£11M** total from NIHR, which was matched by the partner NHS Trusts and the BRCs in the form of funds, existing resources and other contributions. Further funding for the HIC is currently under consideration.

Programme Governance

The Programme is governed by an Oversight Board, comprising the five NHS Trust CEOs; and a Programme Board, comprising BRC Directors and CIOs, with Programme Management teams at each BRC. The Oxford BRC provides the coordinating hub for the programme.

Aims

The aims of the HIC are to:

- establish and maintain catalogued, comparable, comprehensive flows of patient data at each Trust;
- create a governance framework for data sharing and re-use across the Trusts and partner organisations; and
- demonstrate this new capability – the data flows and the governance framework – through the delivery of a number of exemplar research studies, one in each of five specific therapeutic areas.

Research themes

The development of the HIC capability has focussed on five therapeutic areas, chosen for their importance to patients and the NHS and also for the different challenges that they present in terms of data quality and availability. These are:

Theme	Lead Investigator	Lead Partner Site
Acute coronary syndromes	Dr Jamil Mayet	Imperial
Viral hepatitis	Dr Ellie Barnes	Oxford
Critical care	Professor Mervyn Singer	UCLH
Ovarian cancer	Dr James Brenton	Cambridge
Renal Transplantation	Dr Paramit Chowdhury	Guy's and St Thomas'

The success of the programme has led to an increase in scope, with data now also being collected for prostate, breast, colorectal, and lung cancer. Data is also being collected on a wider range of infectious diseases. These new areas will interface with the 100,000 Genomes Project.

Deliverables

The HIC deliverables are:

- Data collection and sharing will be facilitated by a models **(metadata) catalogue**, showing how relevant data is acquired, managed, and made available.
- All sites hold and collect data according to agreed candidate **data models**. The catalogue relates these datasets to models describing local systems and processes.
- Study **datasets** based upon these models will be delivered by each Trust to support the exemplar research projects.
- A **framework to govern data sharing and processing** will be established to support the delivery of these datasets, including a data sharing agreement, allowing for the transfer of de-identified and/or identified information between Trusts.

The Metadata Catalogue

Partner sites have different systems for electronic patient records and little or no data has been shared between them previously. To address this, a single Model (Metadata) Catalogue has been established.

This tool provides a standard way for recording, storing and organising data in a unified format, to create a shared data resource across the partner sites. Context is given to each result or data entry, which makes the data easier to understand and avoids miscommunications and errors being made.

The Shared Datasets

Research leaders from across the five sites have agreed an appropriate, minimum dataset or shared standard for each theme, while the informatics teams at each site have established the clinical data repositories and data flows to extract and link data, and to match these shared standards, ready re-use in support of multi-site research programmes. These are:

i. Acute Coronary Symptoms (ACS) dataset

This dataset contains data from the HIC partners involved in the ACS theme and is stored at Imperial College Healthcare NHS Trust research warehouse. It is an anonymised dataset derived from routinely collected NHS patient data (100k patients and continues to grow) who have received a troponin assay since 2010 (or as far back in time as possible). These are secondary care- local deep episodic patient-level data including: A&E records; pathology tests; drug treatment; cath lab imaging (structural and functional); revascularisation (percutaneous, surgical); echo imaging; transfusion.

ii. Critical care dataset

This dataset contains identifiable data supplied by all participating Trusts (GSTT, CUH, OUH and ICHNT), and stored in the UCL Trusted Research Environment.

It is secondary care - deep episodic patient-level data including: demographic, physiological, laboratory and drug data for all patients admitted to Critical Care and Intensive Care Units from January 2014 onwards (c. 8,000 - 10,000 patients anticipated if all Trusts contribute).

iii. Renal transplantation dataset

This dataset contains de-identified renal transplantation data from the HIC participating centres and is stored within a GSTT research data warehouse. It includes data from patients who have undergone kidney, or pancreas and kidney, transplantation after January 2005, and have been followed up in any of the HIC participating centres. This is a secondary care – local deep episodic patient-level data (currently estimated 6,500 unique patients for the observation period) including information on recipient donor demographics, medical history, transplant details, pathology/biopsy and HLA results, medication and follow up.

iv. Viral hepatitis dataset

Anonymised data from patients (currently 1,200 patients) with a diagnosis of hepatitis B, C, D, E or autoimmune hepatitis, collected from the collaborating sites since 2013, primarily from hospital outpatient clinics. These are secondary care – local deep episodic (per consultant episode) data including admission, treatment, follow up, treatment outcome data.

v. Ovarian cancer

This is a dataset of additional data items to supplement the national cancer data collections (Cancer Outcomes and Services Datasets (COSD) and Systemic Anti-Cancer Therapy (SACT)) to enable research on cancer outcomes in ovarian cancer. This is secondary care – local deep episodic (per surgery, per investigation, per outpatient visit datasets) including: risk factors; surgical resection details; and record of progression events.

Currently, only CUHT has provided data on 400 patients from April 2014; however, the data collection is expected to extend to GSTT and Imperial soon.

vi. Ovarian Cancer – chemotherapy

The dataset has been designed to collect information on all drug treatments with an anti-cancer effect, in all treatment settings, including traditional cytotoxic chemotherapy and all newer agents. Each site obtains its monthly SACT from their cancer reporting team, and run a process to anonymise the patients and select the records with the specified diagnosis range. The collection will be done prospectively in line with the SACT national submission starting with the September 2015 return.

Currently the dataset has 534 chemotherapy records from three Trusts. These are secondary care – local deep episodic (per drug administration of chemotherapy, grouped at regimen and cycle levels) data including: chemotherapy regimen, cycle, drug administration, and treatment end.

Strategic interconnections

- Genomics England – the work of the NIHR HIC involved the development of data models for each clinical theme (mentioned above). In particular, the ovarian cancer model, and learning gathered from a focus on re-use of target data points from national cancer datasets (COSD and SACT) were used as basis for cancer metadata modelling for the NHS 100,000 Genomes Project.
- The University of Oxford/Oxford BRC team who worked on the NIHR HIC data architecture also worked directly with Genomics England (the two programmes share a Chief Technology Officer – Professor Jim Davies), ensuring strategies are aligned and encouraging the consolidation of HIC and the Genomic Medicine Centres (GMCs) approaches within each NHS Trust.
- The metadata catalogue was re-used by the [NIHR Health Data Finder Programme](#). HIC subject matter expertise and toolsets were used in developing the Health Data Finder and ‘onboarding’ meta datasets from NHS Digital, Public Health England and Clinical Practice Research Datalink (CPRD).
- Professor Ellie Barnes (University of Oxford) who leads the HIC Viral Hepatitis theme also leads the MRC Stratified Medicine consortium [STOP-HCV](#), which aims to optimise patient outcomes for Hepatitis C Viral infection. The Oxford team have gathered data on cases from across the five comprehensive BRCs, and the development of a national surveillance system for Hepatitis-C with Public Health England (PHE).

2.2.3. The NIHR Collaboration for Leadership in Applied Health Research and Care

The [Collaborations for Leadership in Applied Health Research and Care](#) (CLAHRCs) undertake high-quality applied research focused on the needs of patients and support the translation of research evidence into practice with the wider NHS and Public Health.

They are collaborative partnerships between a university and the surrounding NHS organisations, focused on improving patient outcomes through the conduct and application of applied health research. They create and embed approaches to research and its dissemination that are specifically designed to take account of the way that health care is increasingly delivered across sectors and a wide geographical area.

The nine CLAHRC pilots were established in October 2008, and have completed their five years of funding. There are now 13 CLAHRCs around the UK (2014-2018) and their geographic coverage aligns with that of the Academic Health Science Network.

2.3. Department of Health and Social Care

2.3.1. Academic Health Science Centres

The Academic Health Science Centres (AHSCs) in England emerged from a commitment by the Department of Health and Social Care to speed up the adoption of medical advances by mainstream clinical practice; and to make the NHS the very best in research and health care. They are a model of healthcare organisation that helps ensure that research breakthroughs lead to direct clinical benefits for the patient by combining world-class research with leading edge clinical services, and excellence in education and training. They also drive economic growth through partnerships with industry, including life sciences companies.

In 2009 a £60M investment was made to establish the first five AHSC in recognition of depth and breadth of clinical and academic excellence. These were renewed for a further five years in April 2014 with the addition of a new Centre. The AHSCs, which are University-NHS Trust Partnerships, are:

1. **University College London Partners AHSC:** University College London, London School of Hygiene and Tropical Medicine, Queen Mary University of London, Barts Health NHS Foundation Trust, Great Ormond Street Hospital for Children NHS Foundation Trust, Moorfields Eye Hospital NHS Foundation Trust, Royal Free London NHS Foundation Trust and University College London Hospitals NHS Foundation Trust
2. **King's Health Partners AHSC:** King's College London, Guy's and St Thomas' NHS Foundation Trust, King's College Hospital NHS Foundation Trust, South London and Maudsley NHS Foundation Trust
3. **Imperial College AHSC:** Imperial College Healthcare NHS Trust, Royal Brompton & Harefield NHS Foundation Trust and Royal Marsden NHS Foundation Trust
4. **Oxford AHSC:** University of Oxford, Oxford Brookes University, Oxford Health NHS Foundation Trust and Oxford University Hospitals NHS Foundation Trust
5. **Cambridge University Health Partners AHSC:** University of Cambridge, Papworth Hospital NHS Foundation Trust, Cambridgeshire and Peterborough NHS Foundation Trust and Cambridge University Hospitals NHS Foundation Trust
6. **Manchester AHSC:** University of Manchester, University Hospital of South Manchester NHS Foundation Trust, the Christie NHS Foundation Trust, Manchester Mental Health and Social Care Trust, Salford Clinical Commissioning Group, Salford Royal NHS Foundation Trust (SRFT)

The AHSCs are corporate entities (federation of partners legally embodied in a company limited by guarantee) with integrated governance and leadership structures that have assumed the role of strategically and operationally managing both healthcare and relevant academic resources.

The six Centres focus on discovery science and the early stages of translational research (T1-T2) within the Trust and usually they have an integrated approach to late stage translation and dissemination of innovation with their local NHS Academic Health Science Network.

The delivery of each of the AHSC's objectives is underpinned by strong clinical informatics platforms/programmes. Details on the specific informatics research activities within each AHSC are provided in the relevant sections of this report describing the wider activities of the Research Organisations.

2.3.2. Genomics England | 100,000 Genomes Project

As part of the strategy to deliver genomic medicine to the NHS, £100M was allocated through the Department of Health and Social Care in England to sequence 100,000 genomes in rare diseases, cancers and infectious diseases from 70,000 patients by the end of 2017. [Genomics England Ltd \(GeL\)](#) was set up to deliver these ambitions and to bring the benefits of personalised medicine to the NHS.

Researchers and clinicians, from both academia and the NHS (*see also section 2.4.2*), are working together to continually analyse data from the 100,000 Genomes Project through the [Genomics England Clinical Interpretation Partnership \(GeCIP\)](#).

The researchers have formed groups or 'domains'. Each [GeCIP domain](#) covers a different disease or topic in the project. The groups are focused on interpreting the data to improve clinical care, but will also undertake complementary research – this includes medical, computational and social research.

Devolved nations

MRC also offered to help the Devolved Governments (Scotland, Wales and Northern Ireland) in their engagement with the 100,000 Genomes Project through GeL. The ambition is to establish a UK-wide partnership to bring together larger assets of genomic data and capability to link to genomic, phenotypic and health record information to offer much greater value and impact than individual collections working in isolation. Plans are driven by the needs and ambitions of the Devolved Governments as well as developing a UK-wide asset and capability.

GeL Data Centre

The Genomics England main Data Centre is based in the **Ark Data Centre** in Corsham Wiltshire.

MRC invested £24M to help establish the Data Centre to deliver the research data management and analysis needs.

2.4. NHS

2.4.1. Academic Health Science Networks

The [Academic Health Science Networks \(AHSN\)](#) aim to deliver a step-change in the way the NHS identifies, evaluates and adopts new technologies and are predicated on partnership working and collaboration between the NHS, academia, the private sector and other external partners, including industry, within a single AHSN context and across AHSNs. This means that they are uniquely placed to identify and spread health innovation at pace and scale; driving the adoption and spread of innovative ideas and technologies across large populations.

There are 15 AHSNs across England, established by NHS England in 2013 as autonomous bio-enterprises and small-medium enterprise in nature. Their impact rests in their ability to bring people, resources and organisations together quickly; delivering benefits that could not be achieved alone. Each AHSN works within its own area to develop projects, programmes and initiatives that reflect the diversity of their local populations and healthcare challenges. However, they all share the following priorities:

- Promoting economic growth
- Diffusing innovation
- Improving patient safety
- Optimising medicine use
- Improving quality and reducing variation
- Putting research into practice
- Collaborating on national priority programmes including:
 - NHS test beds
 - National Innovation Accelerator
 - Supporting New Care Models e.g. NHS Vanguard
 - 100,000 Genomes Project
 - **Data and informatics**

Specific [programmes of activity for 2016/17](#) include the areas of Digital Health; and Data integration/ interoperability.

AHSC vs AHSN

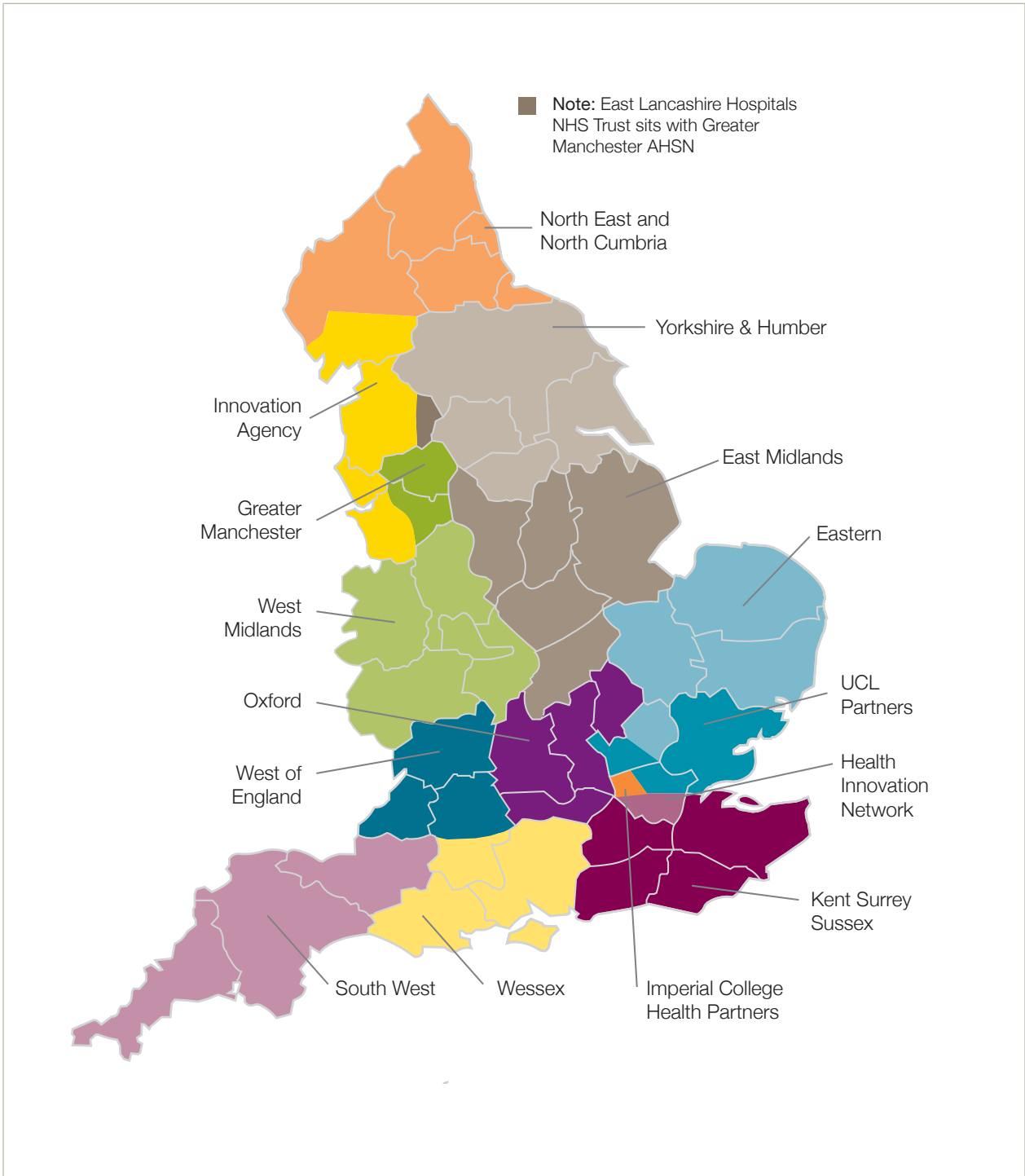
The AHSC focuses on:

- discoveries arising from basic and applied research (T1-T2)
- developing innovation and collaboration with industry
- excellence as a provider of multi-professional education

The AHSN focuses on:

- the diffusion and adoption of discoveries (T2-T4)
- engagement with industry
- new ways of delivering healthcare across a wider population

Figure 4: The AHSN national footprint



2.4.2. NHS Genomic Medicine Centres

Thirteen [Genomic Medicine Centres](#) (GMCs) have been established by NHS England and are pivotal in delivering the 100,000 Genomes Project. The GMCs cover the country delivering an end-to-end genomic medicine pathway. They are enabling the participation of patients and family members with their informed consent; collecting samples to extract DNA; capturing clinical information to inform the interpretation of the genome sequence; and establishing the infrastructure to make genomic medicine a routine part of NHS care.

Figure 5: NHS Genomic Medicine Centres



2.4.3. NHS Global Digital Exemplars⁶

A key component of the [NHS Driving Digital Maturity](#) programme is to create Global Digital Exemplars (GDEs). These are internationally recognised NHS provider delivering exceptional care, efficiently, through the use of world-class digital technology and information. Exemplars will share their learning and experiences to enable other Trusts to follow in their footsteps as quickly and effectively as possible (*Box 3*).

Funding of the GDEs is through the £4.2B 'Paperless 2020' programme announced by Jeremy Hunt in February 2016, overseen by the [National Information Board](#).

Box 3 – NHS England is currently supporting selected digitally advanced mental health and acute trusts, who through funding and international partnership opportunities will become Exemplars over the next two to three and a half years.

Acute Trusts – GDEs

North

1. Alder Hey Children's Hospital NHS Trust
2. City Hospitals Sunderland NHS Foundation Trust
3. Newcastle upon Tyne Hospitals NHS Foundation Trust
4. Royal Liverpool and Broadgreen University Hospitals NHS Trust
5. Salford Royal NHS Foundation Trust
6. Wirral University Teaching Hospital NHS Foundation Trust

Midlands and East

7. Cambridge NHS Foundation Trust
8. University Hospitals Birmingham NHS Foundation Trust
9. Luton & Dunstable University Hospital NHS Trust
10. West Suffolk NHS Foundation Trust

London

11. Royal Free London NHS Foundation Trust
12. Imperial College Healthcare NHS Trust and Chelsea and Westminster Hospital NHS Foundation Trust (joint Exemplar)

South

13. Oxford University Hospitals NHS Foundation Trust
14. Taunton and Somerset NHS Foundation Trust
15. University Hospitals Bristol NHS Foundation Trust
16. University Hospitals Southampton NHS Foundation Trust

Mental Health Trusts – GDEs

1. Berkshire Healthcare NHS Foundation Trust
2. Birmingham and Solihull Mental Health NHS Foundation Trust
3. Mersey Care NHS Foundation Trust
4. Northumberland, Tyne and Wear NHS Foundation Trust
5. Oxford Health NHS Foundation Trust
6. South London and Maudsley NHS Foundation Trust
7. Worcestershire Health and Care NHS Trust

6. Information extracted from the NHS England website

2.5. ESRC's investments in Big Data

The Economic and Social Research Council (ESRC) received £64M in April 2013 to invest in a **Big Data initiative** as part of the wider announcement of capital funding. This covered three distinct areas of activity:

1. **Administrative Data Research Network** – aimed to facilitate researcher access to administrative data routinely collected by Government Departments;
2. **Business Datasafe** – aimed to facilitate researcher access to transactional/ consumer data collected by commercial organisations;
3. **Understanding Populations** – aimed to extend and enhance the ESRC's portfolio of large-scale surveys.

This work was subsequently split into three distinct Phases aimed to deliver the first two areas of activity; and a third covering two further types of data highlighted as important for social science research – social media data and third sector data. The Phases were as follows:

- **Phase 1** focused on the development of the Administrative Data Research Network (ADRN), which provides access to de-identified administrative data collected by government departments for research use. The ADRN awards were made in October 2013. More details are provided in the next section.
- **Phase 2** supported the establishment of centres with a focus on business and local government data. The Centres make data, routinely collected by business and local government organisations, accessible for academics to undertake research that makes a difference: shaping public policies and making business, voluntary bodies and other organisations more effective as well as shaping wider society. The awards were made in March 2014. Further information on each of the investments is provided in *Annex 2*.
- **Phase 3** focused primarily on civil society data and new forms of data. Investment in this area has included: investment in the RCUK Digital Economy Hubs; investment in Civil Society Data Partnerships; New and Emerging Forms of Data Policy Demonstrator Projects; and the Centre for Doctoral Training in New Forms of Data. Further information on each of the investments is provided in *Annex 2*.

2.5.1. Administrative Data Research Network

The Administrative Data Research Network (ADRN) is a UK-wide partnership between universities, government departments and agencies, national statistics authorities, the third sector, funders and researchers. The ADRN is funded until September 2018. ADRN consists of the **Administrative Data Service** which coordinates **four Administrative Data Research Centres (ADRCs)**, all at different locations across the UK:

- **Administrative Data Research Centre England** (Director: Professor Peter Smith): This is led by the University of Southampton with UCL being a key partner. The ESRC has invested in safe settings and technical infrastructure in both Southampton and UCL; **in UCL, the Centre is co-located with Farr@London at 222 Euston Road.**
- **Administrative Data Research Centre Northern Ireland** (Director: Dr Dermot O'Reilly): This is led by Queens University Belfast in partnership with University of Ulster. It is worth highlighting that the service element of the Centre is provided by the Northern Ireland Statistics and Research Agency who provide the data linkage, user service and physical safe access facilities. As a result, ESRC has not made investments in the technical or physical data infrastructures of these two leading universities in Northern Ireland. This is the only ADRC that is not co-located with a Farr Institute Centre.
- **Administrative Data Research Centre Scotland** (Director: Professor Chris Dibben): This is led by the University of Edinburgh in partnership with most universities in Scotland. **The Centre is co-located with Farr@Scotland**, and has developed the technical and physical infrastructure including safe settings in close collaboration with Farr.

- **Administrative Data Research Centre Wales** (Director: Professor David Ford): This is led by the University of Swansea in partnership with University of Cardiff. ESRC invested over £2m in a data science building which also attracted funding from MRC in the context of Farr Institute in Wales. **ADRC and Farr@Swansea share location, staff and other resources.** Safe settings have also been created in the University of Cardiff.

The ADRN helps accredited researchers carry out social and economic research using linked, de-identified administrative data – information which is routinely collected by government organisations. The ADRN makes sure any information which directly identifies people (such as names and addresses) is removed from the data before any research is done. An Approvals Panel examines every research proposal thoroughly. Once they give their approval, the data custodian has to agree to share its data before the research can go ahead. The ADRN provides a secure environment for the researcher to work in. They can't take any data out of this safe setting, and ADRN staff will scrutinise their results before publication to make sure they are relevant to the project and don't disclose any information about individuals.

2.6. EPSRC's health data research investments

The EPSRC supports the development of computational techniques to model, analyse and interpret health and medical data across the breadth of its research funding portfolio, including, but not limited to artificial intelligence technologies, advanced compute infrastructures, information systems and statistics, and applied probability.

Two strategic investments relevant to health data research include:

- **The Alan Turing Institute**

The Alan Turing Institute is a joint venture between EPSRC and five founding universities (Cambridge, Edinburgh, Oxford, UCL and Warwick). The Institute was established in 2015, with a five year £42M government investment, together with a matched £25M investment from the partner universities (£5M per partner). The institute has also formed partnerships with the Lloyds Registry Foundation, Intel, GCHQ and HSBC.

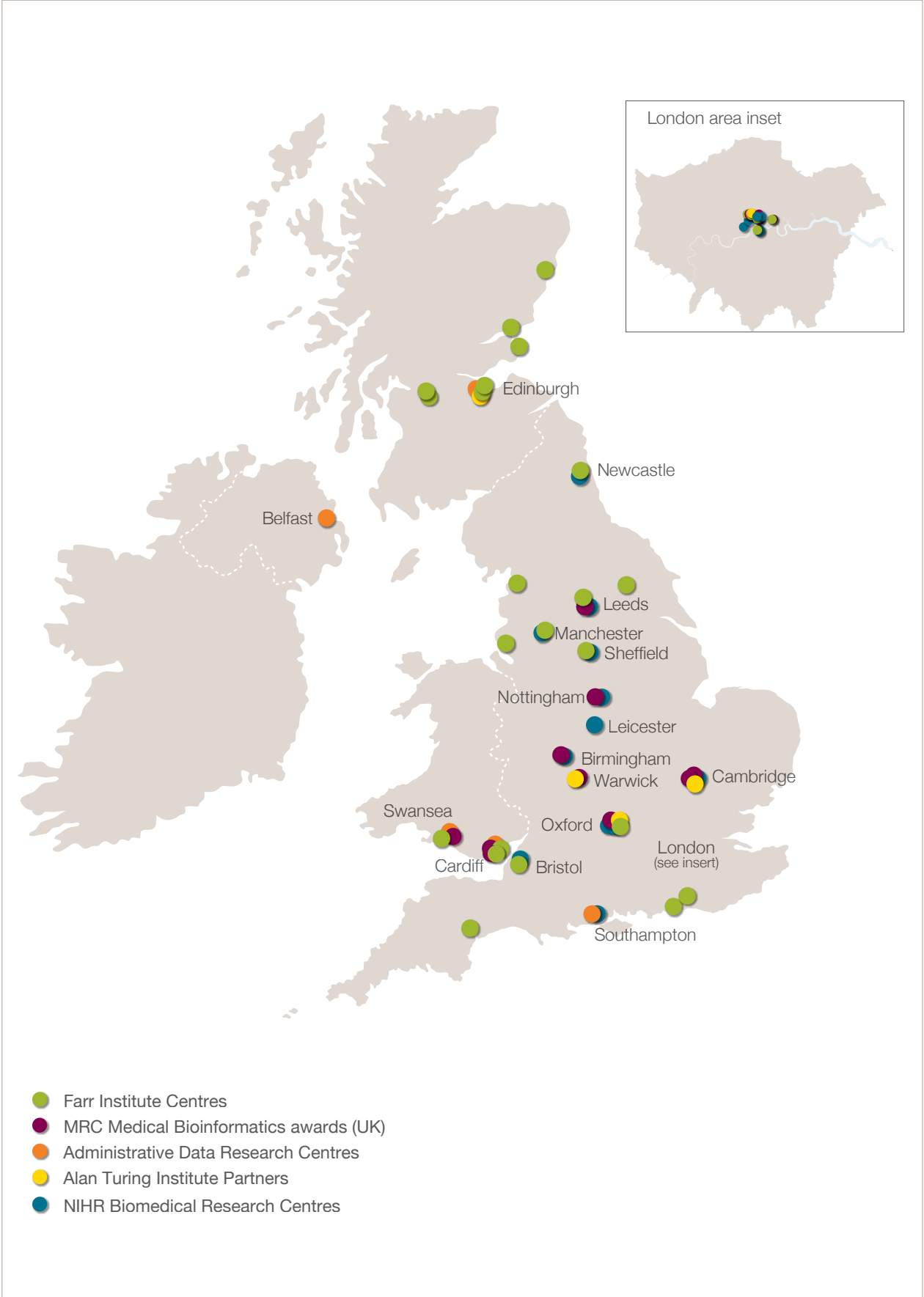
The institutes mission is to undertake data science research at the intersection of computer science, mathematics, statistics and systems engineering; provide technically informed advice to policy makers on the wider implications of algorithms; enable researchers from industry and academia to work together to undertake research with practical applications; and act as a magnet for leaders in academia and industry from around the world to engage with the UK in data science and its applications.

Core research areas for the Institute include data-centric engineering, high-performance computing and cyber-security, to smart cities, health, the economy and data ethics. The institute aims to apply its data science research to real-world problems, working with partners in industry, government and third sector.

- **UK Health Data Analytics Network (UK-HDAN)**

The [UK Health Data Analytics Network \(UK-HDAN\)](#) is a broad community of over 400 health data scientists across the UK that works together to better understand the UK health data analytics landscape. The network has convened several community meetings since being established in January 2016 and drafted a research roadmap, as working document that aims to provide a clear statement of the research challenges and opportunities within health data analytics.

Figure 6: National footprint of key research investments relevant to health data science





CHAPTER 3

KEY INFORMATICS INVESTMENTS BY RESEARCH ORGANISATION

The sections that follow below provide early insights drawn mainly from the Statement of Interest (Sol) submitted from each Research Organisation (ROs) in January 2017. This was in response to a targeted call from the MRC in advance of the call for [HDR UK Phase 1 investment announced in August 2017](#).

In total 26 ROs expressed their interest in becoming a HDR UK Substantive Site and nine ROs expressed their interest in becoming a HDR UK Partnership Site. These have been grouped alphabetically according to their geographic region to make it easier to the reader to navigate through this report – and not for any other reason.

Where possible the information provided within the Sols was cross-checked and supported with further information collected from desk research. Modified “SWOT” analyses were then produced for each Research Organisation and used as tools to extract useful insights.

Please note that this report contains information on the ROs submitting a Substantive Site Sol only

England
East Anglia and Midlands
University of Cambridge
EMBL-European Bioinformatics Institute
Wellcome Trust Sanger Institute
University of Birmingham
University of Leicester
University of Warwick
London
Imperial College London
King's College London
London School of Hygiene and Tropical Medicine
Queen Mary University of London
University College London
North
University of Leeds
University of Liverpool
University of Manchester
University of Newcastle
University of Sheffield
South
University of Oxford
University of Bristol
University of Southampton
University of Surrey
Northern Ireland
Queen's University Belfast
Scotland
University of Dundee
University of Edinburgh
University of Glasgow
Wales
University of Cardiff
Swansea University

GEOGRAPHIC REGION: ENGLAND – EAST ANGLIA AND MIDLANDS

East Anglia

- Separate Statements of Interest (Sols) were received from three Research Organisations (ROs) in this geographic region:

1. **University of Cambridge**
2. **EMBL-European Bioinformatics Institute (EMBL-EBI)**
3. **Wellcome Trust Sanger Institute**

These three ROs propose to submit a joint, strategic proposal to become a Substantive HDR UK Site, if invited at a subsequent stage of the selection process. It is noteworthy that at the Sols stage, these were the only ROs to have proposed such a strategic alliance.

Each of the three individual Sols present a very strong case for becoming Substantive Sites; however, if they join their forces, they will have put together outstanding world-leading, expertise and resources, covering the whole breadth of health data science research from the laboratory to the population to the clinic.

Furthermore, these ROs are situated within the “**Cambridge cluster**”, where large pharmaceutical (e.g. Astra Zeneca, GSK, Pfizer), and technology companies (e.g. ARM, Microsoft) and SMEs (e.g. 1500 high-tech SMEs) have a major presence, offering an exceptional breeding ground for collaborations and further investment.

In addition, the Wellcome Genome Campus has the opportunity to build on the exceptional quality of its current research and service capabilities to become an outstanding centre for scientific, business, cultural and educational activities emerging from genomics and computational biology. As a result, the Campus community is expected to double from 2,000 to at least 4,000 over the next 25 years.

Midlands

- Statements of Interest (Sols) to become a Substantive HDR UK site were received from the following ROs within this geographic region:

1. **University of Birmingham (West Midlands)**
2. **University of Leicester (East Midlands)**
3. **University of Warwick (West Midlands)**

Cross-region Strategic Activities

The Government's agenda for this region

- The Government estimates that the Midlands economy could grow by £34B by 2030 if it matched the predicted growth rate for the UK as a whole. In achieving this, it is also suggested that a further 300,000 jobs could be created by the end of this parliament across the Midlands as a whole.
- **The Midlands Engine Prospectus** is a response to these Government ambitions and sets out the aims for the whole of the Midlands to work more collaboratively to accelerate economic growth and improve productivity across the region. It will also develop into a global brand that allows the Midlands to promote itself in its entirety to the world. To deliver this, it sets out a plan by 11 Local Enterprise Partnerships to encourage economic growth and investment in the Midlands. It focuses on 5 themes: skills, innovation, transport, promoting the plan; and finance for business.

Region-wide health and biomedical informatics activities

- The University of Leicester's multi-disciplinary approach in biomedical informatics – **BINERI (“Biomedical Informatics Network for Education, Research and Industry”)** Leicester – was extended with Nottingham to form BINERI-Nottingham (with joint East Midlands efforts badged as BINERI-EM).
- BINERI – Leicester is also leading a Midlands-wide effort in a **Midlands Bioinformatics Network** aiming to address the challenges and opportunities identified by **Midlands Health Innovation**. The Universities of Leicester, Nottingham and Birmingham now meet regularly to bring added value to each partner's efforts. Specific focus is on four topics: (i) catalogue and make discoverable existing health-relevant datasets and bio-resources; (ii) enable optimal and equitable access to health relevant datasets for research purposes, across organisations; (iii) harmonise consent approaches that enable patients to unambiguously express their desire to have their health data better used; and (iv) establish a platform for a Midlands training network in health and biomedical informatics.
- The **West Midlands Informatics Network** is a group of health informatics professionals aiming to improve health service, by collating relevant information and providing guidance on health informatics solutions for the benefit of healthcare organisations in West Midlands. The Network is looking to support educational initiatives that will enhance health informatics knowledge among healthcare staff and research programmes that will transform patient care. It is also facilitating forums through social media, website, workshops and conferences around priority areas in health informatics.
- The **HPC Midlands Plus** (£3.2M) regional centre was created in January 2017 with investment from EPSRC and contributions from the seven partners: Loughborough University (Lead Site), Aston University, University of Birmingham, University of Leicester, University of Nottingham, University of Warwick, and Queen Mary, University of London. It is currently installing a 0.5Pflop/s service at Loughborough University, of which 15% will be free for use by non-consortium members.

Other region-wide healthcare research related activities

- The **East Midlands Academic Health Science Network (EM-AHSN)** brings together the NHS, universities, industry, third sector, patients and social care to transform the health of the 4.6M East Midlands residents and stimulate wealth creation. EM-AHSN runs award-winning training programmes for Information Analysts, Clinicians and Managers. Current analytics project directly impacting patient care use Big Data analysis to predict re-admissions risk, patient's length of stay to assist capacity planning, and emergency department attendances.

- The **West Midlands AHSN** covers a population of 5.7M, made up of a diverse socio-economic mix. Its aim is to lead, catalyse and drive co-operation, collaboration and productivity between academia, industry, health and care providers and commissioners, and citizens, and accelerate the adoption of innovation to generate continuous improvement in the region's health and wealth. Its major priorities are: manage long-term conditions; promote wellness and prevention of illness; promote the use of advanced diagnostics, genomics and precision medicine; and promote a region-wide approach to the effective management of mental health crisis, reducing pressure on the system and improving outcomes.
- The **NIHR East Midlands Collaboration for Leadership in Applied Health Research and Care (CLAHRC-EM)** is based at the University of Leicester. The CLAHRC-EM coordinates numerous University, NHS, local authority, industry and public partners from across the East Midlands towards improved patient care. The focus is on cardio-metabolic diseases in multi-ethnic populations, integrating primary and secondary local healthcare data with environmental data, to characterise the disorder and optimise treatment. They additionally conduct studies using large data such as CPRD and HES data, and perform interventional studies at practice level with large data capture being conducted remotely across the east Midlands.
- The **West Midlands CLAHRC** is based at the University of Warwick and coordinates numerous University, NHS, local authority, industry and public partners from across West Midlands towards improved patient care. Its focus includes: maternity and child health; youth mental health; prevention and detection; chronic disease; and research methods.

University of Cambridge

Top-level insights

- The University of Cambridge is situated within a powerful cluster of closely-collaborating organisations with expertise in health and biomedical informatics research including: the Cambridge University Hospitals NHS Trusts; the world-leading Sanger and European Bioinformatics Institutes (EMBL-EBI); and the Cambridge Cluster of pharmaceutical and technology companies.
- In addition to the access to world-leading, interdisciplinary, quantitative expertise from across the Cambridge area, the University of Cambridge has also access to significant computational capability including: Peta-5, one of six national Tier-2 HPC facilities for data intensive computation which offers unique capability to test novel data analytics at scale.
- Though its partnership with the Cambridge University Hospitals NHS Foundation Trust, they aim to address major informatics research challenges, with their major focus on the eHospital programme, which led UK's largest ever deployment of an EHR system serving a catchment area of 5M individuals; and the development of a research interface which enables a wide variety of research studies.
- The University of Cambridge has extensive track record of partnerships with industry (e.g. with Astra Zeneca, Microsoft and other large companies and SMEs in the “Cambridge Cluster”); and the NHS (e.g. through the Cambridge Partners Academic Health Science Centre and the East Anglia Academic Health Science Network); and interconnections with other strategic health data-related investments.
- The University of Cambridge has not been supported through the Farr Institute or MRC Medical Bioinformatics awards; hence, the extent of its links/collaborations with these investments is unclear.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The **Cambridge Big Data Initiative**, is a University-wide strategic initiative launched in 2012 to develop a clearly identifiable critical mass of expertise in data science, mobilising expertise in the fundamental technologies of data science and key disciplines needed to make data science work (legal, ethical, societal and economic research).
- The **EPSRC Centre for Mathematical and Statistical Analysis of Multimodal Clinical Imaging** at the University of Cambridge is a collaboration between mathematics, engineering, physics and biomedical scientists and clinicians. It aims to achieve synergies between applied mathematics and statistics through the focus on the analysis of clinical imaging, particularly that arising in neurological, cardiovascular and oncology imaging.
- The **MRC Biostatistics Unit** is a major centre for research, training and knowledge transfer, with a mission ‘to advance biomedical science and human health through the development, application and dissemination of statistical methods’. Its current and recent research on evidence synthesis policy evaluation, new models and trial designs, and association between genetic information in health, has direct impact and influence on clinical practice and public health.

- The **Cambridge Computational Biology Institute** brings together the unique strengths of Cambridge in medicine, biology, mathematics, computer science, physics and engineering. Its aim is to promote excellence in research and teaching and to provide a focus in the Cambridge region for computational biology interpreted broadly.
- The **NIHR Cambridge BRC** (see section further down on 'NHS links') is also one of five **NIHR HIC** sites (see also section 2.2.2) with a focus on cancer data collection to supplement national cancer datasets
- The University of Cambridge is one of the founder members of the **Alan Turing Institute** (ATI). Cambridge academics lead several engagements within ATI's health stream, for example: combining statistics and machine learning with EHR data and wearable devices.

Knowledge/Expertise

- The University of Cambridge has experience in: establishing and leading major data-rich consortia (see next section); developing and applying data science methods and data driven approaches to research; embedding large-scale research into routine NHS systems; and collaborating with industry around the themes of imaging, therapeutic target validation, and stratified trials.

Data

Cambridge coordinates multiple important national and international studies that include:

- Since 2012, Cambridge has provided national leadership for the recall-able **NIHR BioResource** (which builds on the success of the regionally-focused **Cambridge BioResource**). By 2017, the NIHR BioResource included ~60,000 volunteers who provided information on health, lifestyle and biological factors, and consented to be recalled for biomedical research studies according to their genotype or phenotype. High-density gene arrays and metabolomics profiling is in process to facilitate recall. From 2017, Cambridge will lead efforts for a major expansion of the NIHR BioResource, which will aim to recruit ~200,000 recall-able participants with linkages to e-health records. Leadership of such national efforts complements Cambridge-led efforts in building regionally-focussed recallable bioresources, such as the 12,500-participant Fenland Study.
- Cambridge University Hospitals eHospital system provides access to **EHR** to enable a wide variety of new research studies.
- **NHS Blood and Transplant Haemovigilance Platform**, which has linkage of 2M blood donors with EHR.
- **Multi-omics population cohorts with EHR linkages:** INTERVAL (50K participants from 25 centres across England); EPIC-Norfolk (25K participants in East Anglia to study healthy ageing); Cambridge BioResource (14K recall-able participants); Fenland (12.5K participants with serial detailed metabolic assessments).
- Cohort data linked to extensive geospatial information to permit the integration of individual and wider environmental determinants of key health-related behaviours.
- **Strategic contributions to national cohorts:** **UK Biobank** (*Danesh* is a member of the Steering Committee and chair of the Outcomes Working Group; and Wareham leads efforts related to accelerometry (7-day 100Hz waveform in 103,000 participants); and **Dementia Platform UK** (*Richardson* leads biostatistical methods).
- **Pan-European 520K-person EPIC prospective study:** Cambridge leads studies of type 2 diabetes (InterAct), coronary disease (EPIC-Heart), and stroke (EPIC-CVD). An efficient case-cohort design has facilitated measurement of >1M genetic variants and hundreds of soluble biomarkers in a random sub-cohort of ~15K "common controls" who serve as referents for ~40K incident cases.

- **Major international consortia:** the Emerging Risk Factors Collaboration has harmonised individual-participant data on ~2.5M participants in 130 cohorts; Breast Cancer Association Consortium (300K participants), Consortium of Investigators of Modifiers of BRCA 1/2, Global BMI and Mortality Consortium (10M participants).
- **Recall-able population resources in low- and middle-income countries:** PROMIS (cardiometabolic outcomes of 50K participants in Pakistan); BELIEVE (household survey of 50K participants in Bangladesh); BRAVE (cardiometabolic outcomes in 16K participants in Bangladesh); MAVERIK (multi-ethnic study of cardiometabolic outcomes in 5K participants in Malaysia).

Compute

- **Peta-5** is one of six new Tier-2 HPC regional Centres (£20M; EPSRC-funded), which provides access to diverse range of powerful supercomputers. This has been possible through a Cambridge-led national consortium that brings together 15 HPC system and application experts from 10 education institutions for creation of a breakthrough HPC and data analytics capability. Health informatics has been designated as a priority area for use.

Translational and innovation capabilities

Links with industry

- Cambridge is home to one of **Europe's leading pharma clusters**. For example, the University has major strategic partnerships with AstraZeneca (which has recently relocated its £500M global research headquarters to the Cambridge Biomedical Campus), GSK, and several other companies with a major research presence in Cambridge (e.g. Gilead, Pfizer, Takeda).
- The University is also at the heart of the “**Cambridge Cluster**”, Europe's most successful technology cluster of over 1500 small and medium-sized high-tech firms, as well as large technology companies with interests in big data and mobile health, such as ARM and Microsoft.
- Extensive track-record of industry collaboration around themes of imaging, therapeutic target validation, and stratified trials. This has been achieved through bilateral collaborations (e.g. with Microsoft, AstraZeneca, Biogen, GSK, Merck, Novartis, Pfizer) plus multi-institutional partnerships (e.g. EC Innovative Medicines Initiatives such as BigData@Heart).

Links with NHS

Cambridge has a very strong comprehensive translational research pipeline, including informatics research, underpinned by the following investments in partnership with NHS Trusts:

NIHR Biomedical Research Centre

- The **NIHR Cambridge BRC** is a partnership between the University and Cambridge University Hospitals NHS Foundation Trust (£114M; *Director: Dr John Bradley*) and has a major cross-cutting theme in Population and Quantitative Sciences (including clinical informatics, biostatistics and bioinformatics).

This BRC is also a partner in the **NIHR HIC** (*see also section 2.2.2*) leading the ovarian cancer theme (Brenton) aiming to supplement national cancer datasets. It is also informing the cancer strand of the Genomics England national 100,000 Genomes Project.

Academic Health Science Centre

- The **Cambridge University Health Partners Academic Health Science Centre (AHSC)** is a strategic partnership between the University and three local NHS Trusts: Papworth Hospital NHS Foundation Trust; Cambridgeshire and Peterborough NHS Foundation Trust; and Cambridge University Hospitals NHS Foundation Trust. This partnership maximises the potential for effective working with industry, helping to exploit translational funding streams and building capacity for translational health research. Medical informatics features very strongly within the AHSC strategic priorities, which aim to address major informatics research challenges such as the implementation of an EPR (as part of the eHospital programme).

The **eHospital programme** (£200M, 10-year programme) is the largest ever deployment of Trust-wide EPR system in the UK, replacing paper-based records. It allows all members of a patient's care team to access records in real time, whenever and wherever they need to. It comprises an enterprise software solution from Epic, an American company/system, and supported by an entire refresh and upgrading of the IT infrastructure undertaken by Hewlett-Packard (now HPE). The EPR went live on Sunday October 26 2014, and the implementation programme transitioned to 'business as usual' on 1 April 2015.

Cambridge was the first, and so far, the only Trust to deploy the Epic. UCL has just bought Epic which presents opportunities for further collaboration between the two AHSCs.

Academic Health Science Network

- The AHSC is nested within the **East Anglia AHSN**. Through work within the EA AHSN and the East England Ambulance Service Trust, University of Cambridge can link patient data for research supported by their automated data linkage and anonymisation protocols.

Opportunities

- Building on existing **large-scale multi-purpose bio-resources**, the University of Cambridge is pioneering the concept of "living laboratories" for biomedical research by involving participants in translational studies that seek high levels of commitment from participants.
- Cambridge University Hospitals NHS Trust is a NHS **Global Digital Exemplar programme**, one of 13 NHS Trusts to trail blaze NHS's Digital Revolution which will potentially create a breeding ground for innovation in digital health. This has the potential to open new avenues for research.
- In 2016, the University was designated the sole preferred bidder by the Health Foundation, which intends to establish a £41M **Improvement Research Institute** in 2017. Cambridge's bid deploys informatics-based technologies (e.g. crowdsourcing) to identify and evaluate solutions to healthcare challenges.

Potential gaps

- The University of Cambridge has not been supported through the Farr Institute or MRC Medical Bioinformatics awards; and hence, the extent of their links/collaborations with these investments is unclear.
- From the body of evidence collected so far, it seems that the University of Cambridge has not fully developed the areas of digital health and learning health systems research. However, they seem to have the resource and expertise to expand these two areas in the future.

Other relevant capabilities

An overview of further capabilities which potentially align with HDR UK's breadth and strategic interests are presented below.

Research themes

Discovery science

Research activities in this area includes studies on: recall by genotype mechanistic studies; therapeutic target discovery/validation; and imaging.

Precision medicine

- **GAUCHERITE** (MRC Stratified Medicine Consortia £2M, *Professor Tim Cox*): Predictive Measures to stratify clinical outcomes in children and adults with Gaucher's disease and responses to specific therapies. It brings together specialist doctors and scientists led by Cambridge University, who are currently examining approximately 85% of all UK patients and stratifying them by the nature of their disease to allow better targeted therapy interventions.

Population/public health research

Research activities in this area includes studies on: early detection, screening and prevention.

Disease areas

- A number of areas with particular focus on: cancer, CVD, lung, diabetes, infectious disease/AMR.

Analytics

- Data science and precision medicine methods/tools including predictive modelling of health outcomes integrating different data sources; accounting for selection biases created by non-random population selection in health records available for analysis; methodology for analysing health records embedded in clinical trials; longitudinal modelling of disease progression, co-morbidities and dynamic treatment processes; developing new ways to carry out complex evidence synthesis and to deal with informative missing data; exploration of heterogeneity in health outcomes with a view to stratification; development of computational tools and algorithms appropriate to the analysis of data on a large scale.
- University of Cambridge investigators are participating in the **EU FP7 InterConnect** project, which seeks to optimise the use of existing data to enable new research into the causes of diabetes and obesity through the development of co-ordinated platforms for undertaking federated analyses to mirror individual participant data meta-analysis without the governance and ethical challenges of physical data sharing.
- University of Cambridge researchers have developed "**PhenoScanner**", which is the first publicly available tool that provides rapid ability to assess associations for millions of genetic variants with a wide range of diseases, physiological traits, risk factors, and other intermediate phenotypes, drawing on a database of >3 billion genotype-phenotype associations. It has already been accessed >10,000 times by >3000 users in >60 countries. The capability exists to extend innovative tools such as this to help anticipate the broader phenotypic consequences of perturbing specific biological pathways.

e-Infrastructure

Compute

- Peta-5 will provide 3 petaflops (PF) of sustained performance derived from 3 heterogeneous compute elements: 1PF Intel X86, 1PF Intel KNL and 1PF NVIDIA Pascal GPU (Peta-1) connected via a Pb/s HPC fabric (Peta-2) to an extreme I/O solid state storage pool (Peta-3), a petascale data analytics (Machine Learning + Hadoop) pool (Peta-4) and a large 15 PB tiered storage solution (Peta-5), all under a single execution environment.

Training

- University has launched internationally recognised **MSc courses** in: Computational Biology, Machine Learning, Genomics Medicine, and Rare Diseases, which are being cross-linked through modular access with longstanding MPhil courses in Epidemiology, Public Health, and Primary Care Research.
- Examples of 4-year **PhD programmes** in quantitative health sciences in Cambridge include: those funded by the MRC (Biostatistics and Epidemiology Units), Wellcome Trust (Mathematical Genomics and Medicine), and British Heart Foundation (Cardiovascular Genomics, Population Health, Biology).
- An example of **post-doctoral training** in bioinformatics is the joint fellowship programme of the NIHR Cambridge BRC and EMBL-EBI.
- The University also runs **joint fellowships** – at doctoral, post-doctoral, and senior levels – with the Alan Turing Institute.
- For **clinical informatics training**, Cambridge provides senior leadership to the Global Digital Exemplar programme, UK Standardisation Committee for Care Information, and the UK Professional Records Standards Body.

Public Involvement/Engagement

- University of Cambridge researchers have developed scalable approaches to consent, re-consent and re-contact.
- They also engage local communities through “citizen science” projects – for example, use of crowdsourcing to classify tumour data in collaboration with Cancer Research UK.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Alan Turing Institute founders are the Universities of: Cambridge, Edinburgh, Oxford, UCL, Warwick
- Wellcome Trust Sanger Institute
- Peta -5 consortium

DH & NHS

- Cambridge University Hospitals NHS Trusts
- NIHR Cambridge BRC and NIHR HIC
- Cambridge University Partners AHSC
- East Anglia AHSN

Industry

- Pharma: AstraZeneca, GSK, Gilead, Pfizer, Takeda, Biogen, Merck, Novartis
- Tech: Microsoft

International collaborations

- EMBL-EBI
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR
- EU FP7 InterConnect
- EU IMI BigData@Heart

EMBL-European Bioinformatics Institute

Top-level insights

- The European Bioinformatics Institute (EMBL-EBI), an intergovernmental treaty organisation, is a world leader in bioinformatics research and service provision with a well-recognised brand. It provides a highly collaborative, interdisciplinary environment in which research and service provision are closely linked, and take place in an international context, creating a unique organisational culture.
- Based at the Wellcome Genome Campus, EMBL-EBI works closely with its institutional partner and neighbour, the Sanger Institute, to collectively develop the Campus's strategy and offer an exceptional intellectual and physical environment for researchers and organisations with activities in genomes and biodata, not found elsewhere in the UK. As a result, the Campus community is expected to double from 2,000 to at least 4,000 over the next 25 years.
- Both EMBL-EBI and the Sanger Institute have strong links to the University of Cambridge and the Cambridge biotech cluster.
- EMBL-EBI has joint grant funding with researchers and institutes in 62 countries; and is: a partner in several of Europe's emerging research infrastructures, as well as the host of the ELIXIR Hub (Europe's distributed infrastructure initiative for life-science information); and a founding member of the Global Alliance for Genomics and Health (GA4GH).
- Its infrastructure investments include one of the world's largest Life Science Data Centres (100PB storage and 50K cores of compute) and the Embassy Cloud which operates as a service. It is also one of the partners of the MRC Medical Bioinformatics eMedLab private cloud consortium.

Strengths, Capabilities and Opportunities

Strengths

Knowledge/Expertise

- EMBL-EBI is a world leader in bioinformatics research and service provision with a well-recognised brand for quality and innovation. Other activities include data standards development and training.
- It has a long track record in the analysis, storage and management of biological big data and critical biomedical reference datasets, as well as a strong tradition of developing globally adopted databases which curate, organise and present genome and clinical data for research, clinical and commercial communities.
- It has well managed processes and experience in managing data resources jointly funded as international hosts, with sites in two or three countries.
- Its activities are supported by a staff of over 620 personnel comprising of biologists, bioinformaticians, computer scientists, biological data curators, engineers and mathematicians.

Data

- EMBL-EBI hosts the world's most comprehensive range of openly available and up-to-date biological and biomolecular data which include over 40 data resources receiving over 27 million requests per day. Its expertise lies in the analysis, storage and management of biological data at scale, including genomics, transcriptomics, proteomics, metabolomics, molecular pharmacology and biological images.

- Through the European Genome-phenome database (EGA), EMBL-EBI will provide access to the genotype data for all 500,000 participants in the UK Biobank cohort.

Compute

- EMBL-EBI manages a large IT infrastructure in three locations: an on-campus production facility; an off-site, highly resilient Tier III Data Centre which processes all external service needs; and a nearby data replication site. Together these sites have over 100 PB of raw storage and 50k computational cores.
- **Embassy Cloud:** EMBL-EBI operates a cloud service that currently has 30 academic and commercial tenants. The cloud service allows collaborating tenants to bring their computational environment and data close to EMBL-EBI public data resources, where they can undertake computational analysis without the large data transfers this would have traditionally required.
- It also is one of the partners of the MRC Medical Bioinformatics **eMedLab** private cloud consortium (see also section 2.1.2).

Translational and innovation capabilities

Links with industry

- EMBL-EBI data resources and analytics are used extensively not only by academic institutions but also by industry R&D departments. The pharmaceutical industry has traditionally been the biggest industry consumer of their bioinformatics services. As a result EMBL-EBI has a subscription-based Industry Programme for global companies that make significant use of its data and resources. Member companies represent most of the top 20 pharmaceutical companies as well as several major agri-food, nutrition and healthcare companies. The programme is unique, providing regular quarterly strategy meetings, expert-level workshops on topics prioritised by the members, webinars and other activities. It also serves as an interface between industry-focussed initiatives at EMBL-EBI and organisations such as the Innovative Medicines Initiative (IMI), the Pistoia Alliance, the Clinical Data Interchange Standards Consortium (CDISC) and others.
- **Open Targets** is a public-private partnership with GSK, Biogen, Takeda, and Wellcome Trust Sanger Institute for pre-competitive target validation. Current research focus is on oncology, immunology and neurodegeneration through an R&D framework that can be applied to all aspects of human disease.
- EMBL-EBI is also involved in publicly funded projects that have strong industry participation, for example the FP7 and H2020 Innovative Medicine Initiatives (IMI) program and other EU-Framework Programme projects.

Opportunities

- Many of EMBL-EBI's existing data resources provide human genetic and disease information for clinical applications, including: the **Ensembl genome browser**; the **European Variation Archive**; and the **European Genome-phenome Archive**; whilst a collaborative effort is in progress to establish a new **biological image repository** for reference images, which will include medically relevant images of cells, tissues, and organs.
- EMBL-EBI has developed its own **medical strategy**, which is focused on building the necessary research and data infrastructures linking biological information and clinical data to enable the maturation of genomic medicine. Strategic goals include developing clinically relevant data resources, and participating in the construction and definition of infrastructures for medical data such as controlled clinical terminologies and integration with genomic data.

- Immediate plans are to increase support for clinical researchers and practicing healthcare professionals with the provision of bioinformatics services to analyse patient genome data, including the further development of tools such as the **Ensembl Variant Effect Predictor**, a powerful toolset for the analysis, annotation, and prioritization of genomic variants in coding and non-coding regions. Via GA4GH, EMBL-EBI is developing resources for clinical and phenotypic data capture, data use and researcher identification systems and genomics knowledge standards.

Potential gaps

- EMBL-EBI bioinformatics research currently not encompass much work with UK's health data controllers (e.g. NHS Digital) and the data they hold (i.e. routinely collected administrative data) but the expertise is there to do so. With regard to data service provision, EMBL-EBI has a mission to provide publicly available data resources which inherently exclude handling of NHS data.

Other relevant capabilities

Research themes

- EMBL-EBI's main focus is on cutting-edge basic bioinformatics research and the provision of data resources in all major molecular domains; dissemination of tools, data and expertise to academia and industry to enable discovery science translational biology.
- Future research priorities range from genomic analysis to systems biology and include sophisticated multidimensional statistical models for genotype to phenotype associations, evolutionary regulatory genomics, single cell genomics, large scale imaging, phylogenetics, infectious disease genomics, evolution of post translational modifications, integrative systems biology, and structural biology.

Disease areas

- EMBL-EBI supports a wide range of disease areas with a strong interest in cancer genomics including clinical impact (e.g. CR-UK support for PanCancer on the Embassy Cloud) and infectious diseases (WT and MRC support for TB CryPTIC Consortium).

Analytics

- EMBL-EBI hosts over 40 in-house and externally developed software tools for analyses or computation on user-supplied data and has specialised expertise in quantitative methods development, data integration methods, pre-competitive drug discovery tools and other domain-specific tools.

e-Infrastructure

Compute

- EMBL-EBI operates: one of the world's largest Life Science Data Centres (100PB storage and 50k cores); and a cloud service, *Embassy Cloud*, with 31k processor cores in the system; 3.2PB total usable storage; and 200-500 number of registered users.
- Dealing with the rapidly increasing wealth of 'reference' image data that can be linked to external resources will become a priority over the next few years and for this it has already installed 5PB of storage which is going to be expanded.

Training

EMBL-EBI's training and capacity building activities include:

- Currently hosts 31 EMBL International PhD programme students
- Supports three innovative postdoctoral programmes and has a Visitors and Scholar Programme
- Offers advanced bioinformatics training to the community and freely available eLearning material
- Plans to expand its support to clinical bioinformatics training

Partnerships

- Significant experience in leading [global and national partnerships](#) that have contributed substantially to biological discovery.

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- MRC eMedLab – partnership between Crick Institute, KCL, LSHTM, QMUL, UCL, EMBL-EBI and WT Sanger Institute
- Charities: BHF, Wellcome Trust and CRUK
- University of Cambridge, Wellcome Sanger Institute

Industry

- Pharma: AstraZeneca, Bayer, Biogen, GSK, Novartis, Pfizer, Roche and others
- OpenTargets partnership with GSK, Biogen, Takeda, and Wellcome Sanger Institute

International collaborations

- EMBL-EBI has joint grant funding with researchers and institutes in 62 countries
- Other collaborations include: ELIXIR, IMI (number of programmes), Pan-Cancer Analysis of Whole Genomes and GA4GH (Ewan Birney is currently leading this initiative)

Wellcome Sanger Institute

Top-level insights

- The Sanger Institute has world-leading expertise in DNA sequencing at scale, and computational expertise to match these, whilst its research focus is on the integration of health informatics and genomic data to drive novel discovery science.
- In addition, it has an excellent track record of providing bioinformatics/compute support services; and engaging in multi-institutional networks and consortia. For example, it is a founding member of the Global Alliance for Genomics and Health (GA4GH) established to enable responsible and effective sharing of genomic and clinical data.
- Based at the Wellcome Genome Campus, the Sanger Institute works closely with its institutional partner and neighbour, EMBL-EBI, to collectively develop the Campus's strategy and offer an exceptional intellectual and physical environment for researchers and organisations with activities in genomes and biodata, not found elsewhere in the UK. As a result the Campus community has an ambitious vision to lead genomics and biodata in the UK over the next 25 years.
- Both EMBL-EBI and the Sanger Institute have strong links to the University of Cambridge and the Cambridge biotech cluster.
- Its activities are supported by one of the world's largest Life Science Data Centres (45PB storage and 30k cores of compute). It is also one of the partners of the MRC Medical Bioinformatics eMedLab private cloud consortium.

Strengths, Capabilities and Opportunities

Strengths

Knowledge/Expertise

- The Sanger Institute is Europe's largest genomics research institute and one of the leading research institutes globally (1100 employees and over 200 informatics scientists; annual budget £120M).
- In addition, it excellent track record of providing bioinformatics/compute support services and training; advising other large biomedical informatics institutes and initiatives on delivery of computational infrastructure. For example, Sanger holds an MRC Medical Bioinformatics award to develop computational research infrastructure in Africa in partnership with the MRC/UVRI Unit in Entebbe (*see also section 2.1.2*).

Data

- The Sanger Institute has a strong tradition of initiating and developing globally adopted databases which curate, organise and present data to the research, clinical and commercial communities, such as **COSMIC** – the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.
- It has established and expanding programmes coupling genomic research to EHR with **DECIPHER** as a powerful and mature example enabling clinicians and researchers to make matches between patients with rare and ultra-rare disorders.

Compute

- The Sanger Institute operates one of the world's largest Life Science Data Centres, with 1000m² of space 45PB of data storage and 30k processing cores.
- The Sanger Institute is one of the partners of the MRC Medical Bioinformatics **eMedLab** private cloud consortium (*see also section 2.1.2*).

Translational and innovation capabilities

Links with industry

- Wellcome Genome Campus partners include Genomics England, spinouts (VHsquare, Congenica, Microbiotica), and external biodata companies attracted here and housed in the Biodata Innovation Centre.
- The Sanger has also multiple collaborations with pharmaceutical industry partners including: Pfizer, AstraZeneca, Bayer, Roche, GSK, Biogen.
- Working relationships with IT infrastructure and service providers such as Microsoft Cloud, Google and Amazon Web services.
- **Open Targets** is a public-private partnership with GSK, Biogen, Takeda, and Wellcome Sanger Institute for pre-competitive target validation. Current research focus is on oncology, immunology and neurodegeneration through an R&D framework that can be applied to all aspects of human disease.

Opportunities

- Sanger's scientific computing is currently undergoing a transformative programme to enable new capabilities in data sharing, federated analysis and scientific reproducibility.

Potential gaps

- Although the Sanger Institute has strong analytic expertise and tools focused on analysing sequencing data, it has not got extensive experience working with UK's health data controllers (e.g. NHS Digital) and the data they hold.

Other relevant capabilities

Research themes

- Human Genetics – focuses on the integration of health informatics and genomic data to drive novel discovery science. These include:
 - Population-scale assessment of genetic risk factors for complex disease
 - Utilising UK EHRs to enable functional genomic studies
 - Identifying the biomedical consequences of human 'knock-outs'
- Cancer, Ageing and Somatic Mutation
- Infection Genomics
- Cellular Genetics
- Human Cell Atlas
- Malaria

Disease areas

- The Sanger Institute has research programmes in a number of disease areas and in particular: complex diseases (mainly cancer and CVD); rare diseases and infectious diseases.

Analytics

- The Sanger Institute has research programmes in a number of disease areas and in particular: complex diseases (mainly cancer, auto-immune disorders); blood cell traits, rare developmental disorders and infectious diseases.

e-Infrastructure

Compute

- In total the HPC compute environment in Sanger's datacentre contains about 45PB of data storage and has 30k processing cores. These are used for sequencing data storage and the secondary compute and storage for research conducted by its scientists.
- Additionally, Sanger has fully resilient Enterprise systems comprising 3PB of storage and 1500 virtual machine instances.
- It also offers automated self-service access to their HPC/private cloud environments through SaS to internal/external customers and provide support services for design and implementation of HPC environments and clusters. Also, it provides hosting services on request. This compute environment includes 9,000 cores and 3PB internal private cloud using Openstack. The environment has 3PB storage and 50TB of memory and is available to test.
- There is access to the JANET academic network with 2x10Gbit links.

Training

In terms of training and capacity building, the Sanger has:

- experience in developing academic models including Joint Faculty, Associate Faculty, Honorary Faculty and International Fellows that enable established scientists from other institutes to actively engage scientifically with the Institute
- advanced courses and scientific conferences team who develop and deliver postgraduate training, meetings and immersive training in biomedical informatics e.g. MPhil in Genomic Medicine for NHS staff
- a wide range of opportunities for Postdoctoral Fellows, and currently have over 100 postdoctoral scientists
- over 80 University of Cambridge-affiliated graduate students.
- various different types of placement including work experience, undergraduate, masters and sandwich year placements.

Public Engagement

- Partners closely with the Connecting Science initiative, which aims to enable everyone to explore genomic science and its impact on research, health and society.

Partnerships

- Significant experience in leading global and national partnerships that have contributed substantially to genomic research.

Key strategic interconnections with health data research and innovation related investments at a glance:

- EMBL-EBI
- University of Cambridge
- Cambridge University Hospitals NHS Trust
- MRC eMedLab – partnership between Crick Institute, KCL, LSHTM, QMUL, UCL, EMBL-EBI and Wellcome Sanger Institute
- MRC/UVRI Uganda Research Unit Medical Bioinformatics award
- Other Genome Centres (particularly the Broad Institute and the Beijing Genomics Institute)
- Wellcome Trust Centres including notably the Centre for Human Genetics (Oxford)
- Global Alliance for Genomics and Health (GA4GH) – Sanger Institute is a founding member
- Other networks and consortia include: ICGC, INTERVAL, 1000 Genomes, UK10K, DDD, MalariaGEN, MAGIC, HipSci.
- OpenTargets partnership with GSK, Biogen, Takeda, and Wellcome Sanger Institute

University of Birmingham

Top-level insights

- The University of Birmingham combines its strengths in biomedical and health informatics research with those of the University Hospital Birmingham NHS Foundation Trust (UHBFT). The two organisations' health informatics capabilities are based around secondary (hospital-based) and tertiary (referral for specialised services) healthcare aligned with a range of academic expertise that can exploit linked opportunities.
- The UHBFT is continuously developing its EHR system – the backbone of which is the Prescribing Information and Communication System (PICS) that has been used for more than ten years: in an organisation with more than 1 million patient-episodes annually. Importantly, introduction of specific interventions via the EHR has been associated with quality improvements such as reduced readmission, rapid access to proven treatments and a reduction in mortality.
- This EHR system also provides significant benefits with respect to access to some of the world's most complete representation of secondary healthcare data. This provides the opportunity to map demographic and complex clinical data onto phenotyping and genomic information, offering a unique opportunity to explore mechanisms of disease and develop personalised medicine through stratification. Electronic systems developed by UHBFT underpin recruitment to and seamless data collection for the West Midlands Genomic Medicine Centre. These systems will deliver significant legacy capacity for data integration, for the purposes of both service delivery and collaborative research.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- Informatics research is mainly focused around the **Birmingham Centre for Health Informatics**, where the University of Birmingham and the UHBFT have a range of co-funded appointments within disciplines such as bioinformatics, clinical analysis, statistics and social sciences.
- In addition, the two organisations have invested more than £20M in the **Centre for Computational Biology** (*Cazier*) and **Clinical Informatics facilities** (*Ball*).
- As part of the MRC Medical Bioinformatics investment into the **Cloud Infrastructure for Microbial Bioinformatics** (CLIMB; £8.5M), the University of Birmingham is one of four distributed sites (Warwick, Birmingham, Swansea and Cardiff). CLIMB provides infrastructure based on OpenStack, with a variety of virtual server infrastructure provided to the research community. In addition to providing traditional data storage and compute infrastructure, this provides a large geo-replicated object store both to support longer-term storage and to underpin collaboration. The project is working on the automated provisioning of genomic software and pipelines to assist the microbial genomics community. CLIMB currently supports over 1,000 microbiological researchers across the UK (see *also section 2.1.2*)
- The University also holds a number of EPSRC-funded projects within the **Centre for Research in Computational Intelligence and Applications** in areas spanning healthcare software, data usage for personalized medicine and healthcare data security.

Knowledge/Expertise

- Birmingham has expertise in computational science, bioinformatics, integrative systems biology and comparative phenomics and clinical informatics.

Data

- One of Birmingham's most important and unique assets is the amount of data collected on the **Prescribing Information and Communication System**, which forms the backbone of their in-house EHR system (e.g 10,000,000 in-patient drug administrations annually). This provides a comprehensive representation of secondary healthcare delivery and outcome.
- Birmingham's Health Improvement Network (THIN) database (*Marshall*) is a large UK general practice database which contains anonymised longitudinal patient records from over 500 practices (about 6% of the population).

Translational and innovation capabilities

Links with industry

- The University of Birmingham, the UHBFT and the Birmingham Children's Hospital NHS Foundation Trust are members of Birmingham Health Partners (BHP) bringing clinical, scientific and academic excellence across an integrated medical and life sciences campus. They also work closely with Birmingham Women's Hospital, which houses the West Midlands Regional Genetics Laboratory, the largest NHS genetics laboratory nationally. (*Please note that BHP is not one of the AHSC*).
- BHP have established collaborations with many private sector vendors including: EMC, Orion Healthcare, Intersystems, Tibco, BluePrism, Cisco and Blackberry. This ensures access to innovative technological solutions.

Links with NHS

- Although Birmingham has an **NIHR BRC** (£12M from 2017), its focus is on inflammatory arthritis, inflammatory bowel disease and inflammatory sarcopaenia; it is unclear if there are any informatics-related activities.
- Working with the **West Midlands Academic Health Science Network** (WM-AHSN), they coordinate an integrated pan-Birmingham research strategy, translating basic discovery science into clinical benefits through shared infrastructure, governance and expertise. A key example of success is the **West Midlands NHS Genomic Medicine Centre** (WMGMC), the largest hub of the national 100,000 Genomes Project, uniting all 18 regional NHS Trusts and leading and/or involved with several of the national Genomics England Clinical Interpretation Partnerships (GeCIPs). All WMGMC data and samples are linked through the University of Birmingham's Human Biomaterials Resource Centre.
- UHBFT hosts the NIHR **West Midlands CLAHRC** which supports systematic review and evidence synthesis of the impact associated with digital health technologies.

Opportunities

- Access to the rich primary (THIN) and secondary care (HES/PICS) data sets that are available to Birmingham and their integration with genomic and phenotypic data that Birmingham is able to generate at scale can enable a number of novel research projects and development of innovative analytics solutions.
- UHBFT is part of the **NHS Global Digital Exemplar** programme, one of 13 NHS Trusts to trail blaze NHS's Digital Revolution which will potentially create a breeding ground for innovation in digital health. This could potential open new avenues for research.

Potential gaps

- Birmingham will benefit from developing research methodology in the area of learning health systems capitalising on its rich primary and secondary care datasets.

Other relevant capabilities

Research themes

Discovery Science

- Birmingham's genomic capability is represented at both the academic and clinical levels, and includes: the largest **Clinical Genetics** laboratory in Europe; and the **West Midlands NHS Genomic Medicine Centre**. Areas of particular expertise include cancer genomics, microbial genomics (including leading expertise around nanopore technology), DNA repair and rare diseases.
- An **MRC Clinical Research Infrastructure award** (£7.3M, 2014) has helped to establish the first regional **Phenome Centre** (£8M) linked to the National Phenome Centre at Imperial College London and developing unique expertise in advanced metabolic analysis (one of only 4 such Centres globally) and with a capacity to analyse >30,000 samples each year. Importantly, Birmingham now holds the ability to integrate such data with genome and immunophenotyping data.
- The University of Birmingham received a **MRC Discovery Award**: "Placing Discovery Science at the heart of Big Data" (Moss, £300K, 2015-16) to deliver a range of interdisciplinary early phase basic research, which can be subsequently developed to support the UK ambition to be an international leader within the exploitation of biomedical and health data.
- Immunology and Haematology have been strong disciplines on the Birmingham campus for many years, exemplified by the MRC Centre for Immune Regulation and Cancer Immunology and Immunotherapy Centre.

Precision medicine

- As mentioned above, the MRC Clinical Infrastructure Award supports deep immunophenotyping of patients across the campus, enhancing their ability to connect immunology, genetics and metabolomics for stratified medicine. This major 'omics infrastructure allows them to focus on the molecular study and characterisation of different cell types within immune-mediated inflammatory diseases (e.g. arthritis and liver disease) and blood cancers and the associated molecular changes in the patient. Researchers develop and test the integration of multiple diagnostic methods, including mass cytometry, metabolic phenotyping and analysis of single cells.

Population and Public health

- In partnership with the University of Edinburgh (*Professor Aziz Sheikh*), Harvard University (*Professor David W Bates*) and the University of Nottingham (*Professor Anthony J Avery*), they are involved in a multidisciplinary collaboration (**ePrescribing project**) to inform important national deliberations on the safe, effective and efficient procurement and implementation of electronic prescribing systems into hospitals in England. Birmingham (*Coleman*) is leading the work-package on assessing impact on prescribing safety – studying prescribing indicators judged by experts to have a significant risk of translating into actual patient harm.

Citizen-driven/Digital health

- The **Centre for Patient-Reported Outcomes Research** (*Calvert*) aims to optimize the use of such outcomes in clinical trials and routine care, to improve service delivery, enhance patient care and outcomes and ensure that the patient perspective is at the heart of health research and NHS decision-making.

Learning Health Systems

- The **Quality Outcomes Research Unit** (*Pagano*) has developed a number of research outputs including key publications relating to ‘7 day hospital mortality’ that have influenced policy and NHS practice.

Disease areas

- Birmingham has research programmes on a number of disease areas with particular focus on: cancer, diabetes, immunity and infection.

Analytics

- The **School of Health & Population Sciences** brings together major clinical disciplines from the primary and secondary health care settings and has a methodological focus in the areas of health economics, patient reported outcomes and biostatistics. For example, development of algorithms to identify patients with diabetes in hospital patients who are at high risk of adverse outcomes.
- Researchers in the Centre for Patient-Reported Outcomes Research have developed an **electronic patient reported outcome system (ePRO)**, which allows flexible data capture and real-time monitoring of symptoms as well as early detection of problems and intervention from the clinical team. An example includes the current implementation of real-time collection of ePRO reports from patients with chronic kidney disease to improve the information flow between patients and clinicians.
- UHBFT has created the **‘GENIE’ application**, a bespoke tool for data entry, receipt and curation and transmission of data electronically from local systems. For example, a recent NHS England capital award (£1.4M) will allow the seamless receipt and transmission of radiological image (DICOM) files from patients recruited to the 100,000 Genomes Project. The use of GENIE in the 100,000 Genomes Project extends to Great Ormond Street Hospital, Royal Devon & Exeter and Newcastle Genomics Medicine Centres.
- UHBFT has also strengths in the linkage, interpretation and presentation of national datasets. For example, it has developed the **HED product**, an online benchmarking solution designed for healthcare organisations. This which provides business analytics to 67 organisations across the UK. It uses data sources such as: HES, HPA Data, HRG Tariff, NPSA, ONS Mortality, Workforce and VTE Risk Assessment to allow national benchmarking across a wide range of clinical indicators.

e-Infrastructure

Compute

- Built with a major investment by the University (including £2M for Life Sciences) the **Birmingham Environment for Academic Research cloud** infrastructure provides compute and 5PB of storage to the life sciences community. Its HPC design is based on OpenStack, providing flexible computing environments with both virtual machines and classic cluster infrastructure.
- The **Cloud Infrastructure for Microbial Bioinformatics** (CLIMB) includes a distributed four-site computing system capable of supporting over 1,000 virtual bioinformatics servers. This system consists of hardware with 7,680 cores, 2,304 TB of cross-site replicated storage and 78TB of RAM.

Platforms

- UHBFT holds a unique position as an NHS Trust that has developed, and successfully deployed, optimal clinical systems both locally and into other NHS Trusts. For example, UHBFT has developed the **OPTIMs Outpatient Management System**, which has been delivered to over 40 UK hospitals (by a commercial partner).

- UHBFT has developed the **Prescribing Information and Communication System (PICS)**, which forms the backbone of their EHR system. It is a rules-based prescription support system that can be configured by speciality, allowing real-time drug prescribing, and checks for drug interactions, contra-indications, dose limits and more. In terms of research provision, this places them in a unique position of being able to collect highly detailed datasets that can be used to address very specific research questions.

Training

- The University of Birmingham's existing educational capabilities in informatics research that include an **M.Sc. in Bioinformatics** and an **MSc in Health Research Methods** (with Health Informatics module).
- The Centre for PROs Research provides education through online learning resources, CPD workshops, with a Masters module in outcomes methodology in development.
- Other informatics-relevant courses under development include a massive open online course (MOOC) in digital healthcare; and a postgraduate certificate course on health informatics.

Public Involvement/Engagement

- As part of the UHBFT EHR system, they have developed the '**myHealth@UHB**' **patient portal** which allows patients to access their EHRs. There are currently 14,400 registered users, recruited primarily from chronic disease specialties. There is strong patient-level support for access to these records, and they are involved in co-design through their PPI, clinical and academic leads.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- MRC Medical Bioinformatics CLIMB – award partners include: Birmingham, Warwick, Cardiff and Swansea
- Farr Institute (collaboration on patient outcomes research)
- Imperial College London (collaboration with the Phenome Centre)
- University of Edinburgh University of Nottingham (collaboration on the ePrescribing project)

DH and NHS

- Genomics England/100,000 Genomes Project
- West Midlands GMC and other GMCs
- NIHR Birmingham BRC
- NIHR West Midlands CLAHRC
- West Midlands AHSN

Health data providers

- Public Health England
- IMS Health
- TTP

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:**Industry**

- Pharma: Pfizer, Janssen, Takeda, Merck, ABPI
- Tech: MSDi, EMC, Orion Healthcare, Intersystems, Tibco, BluePrism, Cisco & Blackberry

International

- Harvard University
- University of Sidney
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR

University of Leicester

Top-level insights

- University of Leicester has a well established strategic partnership with the University Hospitals of Leicester NHS Trust as exemplified by the joint activities in the Leicester Precision Medicine Institute and the Biomedical Informatics Network for Education, Research and Industry (BINERI-Leicester).
- The University of Leicester is leading a Midlands-wide effort in a Midlands Bioinformatics Network and is working with other regional institutions and organisation through the Midlands Engine and new Enterprise Zones to drive cross-cutting initiatives and advance the regional economy.
- Together with partners at Loughborough University, University of Leicester have a MRC-EP SRC Pathology Node in Breathomics (EMBER), with plans now under development for a new cross-cutting Centre in Respiratory and Environmental Health.
- Leicester has recently launched an Institute for Space and Earth Observation, which undertakes world-leading EPSRC- and ESRC-funded research in pollution, weather and environmental data collection and analysis, coupled to nationally leading HPC capabilities and technological and analytical methods development.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The **Department of Informatics** (formerly the Department of Computer Science) research focuses on five broad themes: (i) Algorithms: Complexity and Engineering; (ii) Foundations of Computing; (iii) Interaction Design and Evaluation of Socio-technical Systems; (iv) Software Modelling; and (v) Evolution, Validation and Verification. There is also an active working group on Knowledge Discovery and Machine Learning.
- The **Leicester Institute for Space and Earth Observation (EO)** is spread across a number of Departments within the College of Science and Engineering: Physics and Astronomy, Chemistry, Computer Science, Geography, Geology, Engineering and Mathematics. The focus is on missions, instruments, data and innovation. Expected outcomes will bring science leadership alongside engineering capability, and develop expertise on data analysis, data exploitation and leading technology that can be applied both within space research and outside in other areas – for example, medical devices and diagnosis (already underway) – using their advanced HPC capabilities.
- The University of Leicester is one of the partners on the **MRC CLIMB** Medical Bioinformatics investment. This cloud infrastructure based on OpenStack, provides a variety of virtual servers to the research community. In addition to offering traditional data storage and compute infrastructure, it provides a large geo-replicated object store both to support longer-term storage and to underpin collaboration. The project is working on the automated provisioning of genomic software and pipelines to assist the microbial genomics community. CLIMB currently supports over 1,000 microbiological researchers across the UK (see also section 2.1.2)

Knowledge/Expertise

- The University of Leicester has established a multi-disciplinary approach to combining expertise across biology, engineering, statistical methods and clinical practice. For example, **BINERI-Leicester** is a synergistic network that has been active for more than five years. Members comprise senior staff that bridge university and the NHS, including informatics researchers, Leicester's BRC, computer scientists (Department of Informatics), IT Services teams, bioinformatics and analytical services, social scientists, physicists (earth observation and environmental researchers), and mathematicians (e.g. a network on 'Data and Modelling in Medicine' – DaMM).
- The University of Leicester has expertise in delivering **data standards** suitable for connecting and enabling federated querying of resources. For example, they lead a GA4GH task team that produced the Automatable Discovery and Access Matrix (ADA-M) standards for computer-readable representation of consent.
- Leicester staff also co-Chair the IT committee of **BBMRI.uk**, and formally provide IT expertise to the **UKCRC Tissue Directory and Coordination Centre** (led by the University of Nottingham).

Data

- The **Extended Cohort for E-Health, Environment and DNA (EXCEED)** project has access to data on ~8,000 adults, enriched for chronic lung disease, but applicable to other areas;
- The **Genetics and Vascular Health Check (GENVASC)** study combines clinical (recruitment and electronic data transfer from >100 GP surgeries – 22,968 patients) and genetic data to predict coronary artery disease.

Compute

- The University of Leicester IT support and HPC capacity are nationally recognised, as evidenced by hosting one of four nodes of the BEIS/STFC 'DiRAC' national HPC Facility, and imminent Directorship of DiRAC.

Translational and innovation capabilities

Links with industry

- The University of Leicester has three main vehicles for industry engagement:
 - **Medilink**, a life science business support and investor network that hosts a 'Digital Health Special Interest Group';
 - the **Midlands Engine Network** that links HEI and business across the Midlands to direct regional investment and growth; and
 - Innovate UK **Knowledge Transfer Network-Med Tech and Health**, facilitating mixed industry and academia workshops.
- The University of Leicester also works with industry on **EU-IMI projects**.

Links with NHS

- The University of Leicester in partnership with the University of Loughborough and University Hospitals of Leicester (UHL) NHS Trust received an NIHR award (£11.6M) in 2017 to establish a **new BRC** in illnesses linked to respiratory conditions, cardiovascular disease, type 2 diabetes, chronic kidney disease and the consequence of inactivity.

- The Leicester region is covered by the **East of England NHS Genomics Medicine Centre**.
- Furthermore, University of Leicester staff are partners in 11 different Genomics England Clinical Interpretation Partnership (GeCIPs), with the Leicester **Precision Medicine Institute's** Director (*Martin Tobin*) leading the cross-cutting **methodological GeCIP** entitled "Machine Learning, Quantitative Methods and Functional Genomics". This has been invited to be at the vanguard of accessing the data embassy ahead of other GeCIPs to help test the system. They also lead the Analytics work of the Musculoskeletal GeCIP.
- The **East Midlands Collaboration for Leadership in Applied Health Research and Care** (CLAHRC-EM) is led by Professor Kamlesh Khunti based at the University of Leicester. The CLAHRC-EM coordinates numerous University, NHS, local authority, industry and public partners from across the East Midlands towards improved patient care. The focus is on cardio-metabolic diseases in multi-ethnic populations, integrating primary and secondary local healthcare data with environmental data to characterise the disorder and optimise treatment. They additionally conduct studies using large data such as CPRD and HES data, and perform interventional studies at practice level with large data capture being conducted remotely across the East Midlands.
- The University of Leicester is actively involved in the **East Midlands AHSN**, which brings together the NHS, universities, industry and social care to transform the health of their 4.6m East Midlands residents and stimulate wealth creation.

Opportunities

- The University of Leicester provides specialist opportunities having: (i) majority Black and Minority Ethnic communities from South Asia and sub-Saharan Africa that provide additional opportunities to expand research with partner organizations in low and middle income countries; and (ii) big data and expertise from their space, earth observation and pollution research.
- In terms of the second opportunity mentioned above, an example would be establishing a national effort to use environment and earth observation data to guide healthcare capacity planning in semi-real-time. Unifying the efforts of the few ROs sufficiently advanced in this topic to contribute would be valuable, merging this with input from groups such as the UK Meteorological Office, Public Health England (their 'Environmental Public Health Surveillance System'), the London School of Hygiene & Tropical Medicine, the University of Cambridge, and the University of Exeter (their 'Medical & Environmental Data Mash-up').

Potential gaps

- The University of Leicester has not developed significant activity in the areas of learning health systems and citizen-driven/digital health research. It also does not seem to have developed any public involvement/engagement activity relevant to health data science.

Other relevant capabilities

Research themes

Discovery Science

- The University of Leicester is leading the first genetic study in UK Biobank (**UK BiLEVE**) that linked genotypic and phenotypic data and returned this to UK Biobank for wider exploitation.

Precision medicine

- **Leicester Precision Medicine Institute** (*Director: Martin Tobin*) unites the research expertise and facilities of the University of Leicester and University Hospitals Leicester NHS Trust with industry to provide new discoveries and medical innovations that impact on the healthcare of Leicester's ethnically diverse population.
- Leicester researchers (Brightling) have developed and validated the utilisation of imaging and biological biomarkers in early phase clinical trials and the management of airway disease including coordination and participation in several multi- centre, multi-national consortia – for example: AirPROM, COPDMAP and, EMBER, EvA, and ATLANTIS.
- **MRC-EPSCRC Molecular Pathology Node in Breathomics** (EMBER; *PI: Chris Brightling*) – University of Leicester along with the University Hospitals Leicester NHS Trust and Loughborough University lead one of six Pathology Nodes. It is unique in generating new data generation in the form of 'Breathomics' (i.e., breath analysis), using technology engineered for the detection of explosives in war zones. This cross-discipline approach is now being applied to respiratory and other disorders. The method requires advanced big data analytics for near-real-time signal interpretation. Methods for this are developed by Leicester's DaMM team in conjunction with mathematicians from Loughborough, simultaneously linking in clinical and 'omics data.
- University of Leicester has expertise on data management and analysis in several large stratified medicine and personalised medicine consortia in the UK and Europe and are now working with QMUL to determine how to achieve added value from uniting datasets and approaches from such projects (data structure level, software level, and discovery/sharing level), starting with MRC Stratified Medicine projects concerned with COPD, asthma, RA, SLE and Psoriasis.

Population and Public health

- University of Leicester researchers (*Draper*) lead national and international research programmes focussing on the variations in outcome for babies, infants and children. This work involves the linkage of national databases with births and deaths data from ONS (and other similar organisations across the UK) as well as HES data. They are also leading data harmonisation efforts (H2020 RECAP – Research in Children and Adults born Preterm) which aims to link together all preterm birth cohort databases across Europe and with other international partners to enhance the power of the individual studies to enhance the current evidence base for preterm infants.
- Projects connecting health-related datasets with space and earth observation data include:
 - NERC "EPI" project with Cerner to predict hospital admissions, soon modelling 15 years of NHS admissions data; and
 - MRC "ERICA" project (with Bristol, Leicester and Imperial) making environmental data usefully discoverable.

Citizen-driven/Digital health

- The University of Leicester together with Medilink, a life science business support and investor network, run a Digital Health Special Interest Group.

Disease areas

- The University of Leicester has research programmes on a number of disease areas with particular focus on: respiratory conditions, cardiovascular disease, type 2 diabetes, chronic kidney disease.

Analytics

- The University of Leicester has extensive expertise in biostatistics method development – for example, creating ways to handle time-to-event data (risk prediction, life expectancy etc), joint modelling that links longitudinal biomarker data/trajectories to clinical events, and approaches to multi-state modelling for complex disease profiles.
- Researchers at the University are leading the development and implementation of **BRISKit**, an integrated open-source biomedical research software solution (Jisc/HEFCE and user funded), across a range of public health, environmental health and multi-disciplinary research challenges. It is also currently deployed in support of 100k Genomes Programme in UHL NHS Trust.

e-Infrastructure

- The University of Leicester provides a substantial **shared HPC resource** to all researchers, comprising: 4,760 Intel Xeon x86 (Broadwell) CPU cores delivering 180Tflop/s computation; 170 compute nodes each with 28 CPU cores, 128GB RAM and Infiniband FDR (56Gb/s) network; 1.3PB of high performance parallel file system (Lustre); with supplementation by 12 compute nodes, each with 28 CPU cores and Infiniband FDR network.
- The University of Leicester is connected to **JANET** via resilient 10Gb/s links, with most connections to their departments on UHL sites also being in many cases 10gb. A scalable, shared storage resource of ~1PB, is provided free to all researchers.
- The University of Leicester is one of only four UK HEIs hosting HPC clusters for the **DiRAC** national HPC Facility – a multi- architecture Tier-1/Tier-2 service funded by BEIS and STFC. Within DiRAC, it supports a 4,500-core HPC system (growing to 7,500 in April 2017) and from July 2017, Mark Wilkinson (Leicester) will be the national Director of the facility.
- University of Leicester is also a member of **HPC Midlands+** consortium which is currently installing a 0.5Pflop/s service at Loughborough University, of which 15% will be free for use by non-consortium members.

Training

- The University of Leicester runs **MSc** courses in: Bioinformatics; and Medical Statistics; and funds multi-disciplinary PhDs across its Biology/Medical and Science/Engineering Colleges.
- It also runs a 'BBASH' bioinformatics training and service facility for researchers; offers intercalated year projects in Data Science for medics; provides Masterclasses in Bioinformatics for PhD students, include a Bioinformatics module as part of biology degrees; has allocated strategic research funds to give on-the-job bioinformatics training to biology postdocs.
- The University of Leicester is now exploring the option of a **Biomedical Informatics degree course** (connected to a dual-degree programme with the Chinese University of Hong Kong, at their Shenzhen Campus).
- A **Leicester Health and Biomedical Fellowship Academy** provides match-funding for fellowships and associated infrastructure, with partnerships across the Midlands (e.g., the Midlands Bioinformatics Network).
- The University of Leicester runs a **Research Software Engineering Team** to provide all researchers with access to expertise in software development and optimisation. The group's initial focus on HPC projects is now expanding to other informatics areas, whilst also providing regular training workshops to promote best-practice in software development across the University.

Public Involvement/Engagement

- The University of Leicester's College of Medicine, Biological Sciences and Psychology has long recognised the importance of engaging with the public, not only to enhance and inform their research and teaching, but also to strive to make a real and positive impact on the wider community.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- MRC Medical Bioinformatics CLIMB – encompasses the following partners: University of Warwick, Cardiff University, Swansea University, University of Birmingham, University of Bath, University of Leicester and The Quadram Institute (a BBSRC research institute)
- EPSRC, ESRC, NERC
- Universities: Birmingham, Nottingham, Loughborough, UCL, Imperial, Bristol

NHS and other government organisations

- NIHR Leicester BRC
- University Hospitals of Leicester (UHL) NHS Trust
- East of England NHS Genomics Medicine Centre
- Genomics England – GeCIPs
- EM-AHSN
- EM-CLAHRC

Data providers

- CPRD
- NHS Digital
- ONS

Industry

- Medilink

International

- EU BBMRI
- EU IMI
- GA4GH
- IRDiRC
- Chinese University of Hong Kong

University of Warwick

Top-level insights

- The University of Warwick has strengths in developing novel analytical methods and translating these to address practical challenges in health as evidenced by a number of investments including the MRC Medical Bioinformatics award into the Cloud Infrastructure for Microbial Bioinformatics (CLIMB). Other major collaborations in data science include the Alan Turing Institute, with the University of Warwick being one of the five constituent universities.
- Warwick's Institute for Digital Healthcare is a collaboration between the Warwick Manufacturing Group, the Warwick Medical School and the NHS developing solutions needed to address modern challenges across all aspects of healthcare. The Institute collaborates with Farr Institute partners. Other interdisciplinary University investments relevant to health data science include: the Warwick Data Science Institute; and the Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research.
- The University of Warwick has research and implementation partnerships at all levels in the local health ecosystem which reaches 5M people in Coventry and Warwickshire, and 10M across the West Midlands, including: West Midlands AHSN; and CLARHC. In addition, major resources for health and bioinformatics research will be further developed as part of the Midlands Engine Prospectus.
- Although, the University of Warwick was not part of the Farr Institute, it has been collaborating with its partners.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

The University of Warwick's major inter-disciplinary investments in the areas of biomedical research and population health with relevance to informatics include:

- The **Warwick Data Science Institute** (WDSI, Director: *Professor David Firth*) brings together not only the strengths of the Departments of Computer Science, Mathematics and Statistics, but also the wider community both at Warwick (researchers from many departments, with interests ranging widely from genomics to business analytics) and internationally. Its primary research focus includes the development of novel mathematical, statistical and computational approaches to the acquisition, management and analysis of big data.
- The **Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research** (SBIDER) brings together a critical mass of interdisciplinary scientists from mathematicians, statisticians and computer scientists to biologists, biomedical researchers, and clinicians.
- The **Institute for Digital Healthcare** is a partnership between Warwick Manufacturing Group, Warwick Medical School and the NHS which aims to develop the digital solutions needed to address modern challenges across all aspect of healthcare.
- The **Alan Turing Institute** (ATI) for data science, where University of Warwick is one of the five constituent universities together with Cambridge, Edinburgh, Oxford and UCL.

- As part of the MRC Medical Bioinformatics investment into the *Cloud Infrastructure for Microbial Bioinformatics* (CLIMB; £8.5M; see also section 2.1.2), Warwick is one of four distributed sites (Warwick, Birmingham, Swansea and Cardiff) providing infrastructure based on OpenStack, with a variety of virtual server infrastructure provided to the research community. In addition to providing traditional data storage and compute infrastructure, this provides a large geo-replicated object store both to support longer-term storage and to underpin collaboration. The project is working on the automated provisioning of genomic software and pipelines to assist the microbial genomics community. CLIMB currently supports over 1,000 microbiological researchers across the UK (see also section 2.1.2)

Knowledge/Expertise

- University of Warwick's strengths lie in the development of novel analytical methods and the translation of these methods to address practical challenges in health. Thus, it has established a track-record of providing research-driven advice, and data solutions, to health providers ranging from local hospital Trusts to national and global organisations.
- The Warwick Medical School's Microbiology and Infection Unit has internationally recognised expertise in pathogen genomics and meta-genome bioinformatics, including: the storing and processing of large-scale next-generation sequence data; and the development of novel tools and databases for its analysis in public health contexts.

Data

- The **Enterobase** is one of the most comprehensive curated pathogen genome database, containing >100,000 assembled, genotyped and annotated genomes from Salmonella, Escherichia, Yersinia Moraxella and Clostridiodes difficile. It provides the only website that contains all the assembled genomes from the short reads published by Public Health England, Sanger Institute and FDA. It is used internationally by over 700 individuals and diagnostic reference laboratories, to identify close relatives of epidemic strains, bridging the gap between research and public health.

Translational and innovation capabilities

Links with industry

- **The Warwick Manufacturing Group** has a renowned track record for translating academic excellence into industrially relevant technology, working with large multinational companies (such as Jaguar-Land Rover, IBM and Nikon).

Links with NHS

- The University of Warwick is a core partner in the **West Midlands Academic Health Science Network (AHSN)** and is leading its **Health Informatics Network** which supports the adoption of technology solutions to transform patient care.
- It is also the academic lead for the **West Midlands Collaboration for Leadership in Applied Health Research and Care (CLAHRC)** which works across health and social care organisations in the region to create effective partnerships and includes a dedicated theme on "**Implementation and Organisation and Research Methods**".

- The University of Warwick has taken a leading role in the development of the **Coventry and Warwickshire NHS Sustainability & Transformation Plan** and its ambitious plan to implement an integrated EPR linking primary to acute care. Local collaborations include: University Hospitals Coventry and Warwickshire (UHCW) digital pathology centre of excellence; University Hospitals Birmingham (UHB) computational models for childhood brain tumour characterisation; Heart of England NHS Trust; Public Health England's real-time syndromic surveillance team (ReSST) in Birmingham; and local Public Health Microbiologists.

Opportunities

- The University of Warwick ranks among the top 10 UK Universities (based on REF2014) in computer science and informatics. Its established expertise in theoretical maths, statistics and computer science as well as fundamental biology and applied health research can catalyse HDR UK activities by disseminating existing, but also developing new cutting-edge analysis tools and methodologies.
- In their Sol, the University of Warwick proposes to extend its component of the MRC Biomedical Informatics investment in **CLIMB** through interactions with HDR UK in three ways:
 1. implementing novel statistics and bioinformatics;
 2. associated training to the UK medical community; and
 3. extending the CLIMB remit to new areas of health-related bioinformatics.

Potential gaps

- Although the University of Warwick has research and expertise relevant to stratified medicine, it is not leading on any major initiatives/programmatic activities

Other relevant capabilities

Research themes

Discovery Science

- **Pathogenomics**, supported by funding from the Wellcome Trust, is a major effort to sequence 1000s of ancient DNA samples from humans and their domesticated animals as well as 10,000s of modern bacterial genomes. The data will be stored and analysed directly within Enterobase, which will include modern pipelines for identifying genomes within metagenomic mixtures from multiple organisms.
- **Systems Biology and Bioinformatics** research within SBIDER brings together strength in mathematical and statistical analysis with a deep understanding of biological systems and technologies, involving collaboration with many leading biological / biomedical labs around the world.
- **Image Analysis**, with particular strengths in analysis of brain and digital pathology images and dynamic imaging of live cells and tissues. This research aims at linking the emerging image data with other modalities such as underlying genetic and clinical information, and the aim of inferring process from image.
- **Bioimaging**, on-going work includes a substantial part of the NIH-funded **Human Connectome project**; and mining of large collections of tissue images for patterns that carry predictive power for disease outcome and survival in collaboration with multiple healthcare providers.

Precision medicine

- In stratified/precision medicine, University of Warwick researchers apply approaches that combine cutting-edge statistical methods with mechanistic (predictive) understanding of human physiology. For example, in chronotherapy for cancer patients with funding from CR-UK, BBSRC and EU.

Population and Public health

- **Epidemiology and Infection**, particularly mechanistic model formulation, novel methodologies, prediction, analysis and inference, and strengths in immune-system modelling, vaccine preventable infections and neglected tropical diseases. This quantitative research is supported by expertise across campus in AMR, public health and health economics, field epidemiology and health-related bioinformatics.
- The **Populations, Evidence & Technologies** group brings together researchers with interests in population research, e-health and health services, with an emphasis on health technology assessment.
- The **Communicable Disease Control Evidence & Epidemiology** team is co-funded by UoW, Public Health England and Health Education England. Current work includes: projects linking human and animal disease datasets to inform on zoonoses; the use of routinely collected datasets to monitor the impact of immunisation programmes; and the value of information to guide disease surveillance.
- The **Health Economics group** is providing detailed health economic advice to the Department of the Health and the Joint Committee on Vaccination and Immunisation (JCVI). This 5-year DH-funded support role combines modern epidemiological modelling matched to the latest health datasets with health-economic forecasts to predict the costs and benefits associated with vaccination policies including meningococcal B, pertussis, HPV and pneumococcal vaccination.
- Funded by the leading pregnancy charity, Tommy's, University researchers are working with diverse clinical teams to investigate the causes of early miscarriage, using data from over 6M patients.

Citizen-driven/Digital health

- The Institute for Digital Healthcare is involved in both national and international projects including, the **C3-Cloud project** (€5M, EU H2020; Arvanitis), which seeks to transform care for patients with multi-morbid chronic conditions by establishing a new ICT infrastructure to enable integrated care.

Learning health systems

- The Institute for Digital Healthcare has also participated in the **TRANSFoRm** project (€9.7M EU FP7, 2010-2015; Arvanitis) leading the User and Software Services Workpackage. The project developed, piloted and evaluated a core information architecture for the Learning Health System (LHS) in Europe.

Disease areas

- The University of Warwick has research programmes on a number of disease areas including: cancer, communicable diseases, infection, as well as neuropsychiatric and neurodegenerative disorders.

Analytics

- The **Communicable Disease Control Evidence & Epidemiology** team undertakes novel methodological development including: improved use of routine data for public health; and ways to combine statutory and research data to better inform health protection interventions.
- **Warwick Evidence** (within the Populations, Evidence and Technologies group) is developing novel statistical techniques to analyse combined datasets (trials and clinical practice) for nationally commissioned evidence synthesis – for example, leading work for the NICE and the National Screening Committee.
- There are also very strong links between the **Zeeman Institute** and PHE's Centre for Emergency Preparedness and Response, who are interested in developing novel computational techniques based on mechanistic models that can inform about early epidemic dynamics.

- Receiving funding from the NIH, University of Warwick researchers (Nichols) are developing Bayesian spatial point process modelling methods to analyse neuroimaging coordinate-level data, binary imaging data and voxel-level data.

e-Infrastructure

Compute

- Scientific Computing RTP: Tinis2 cluster is based on Intel E5-2680 v4 (Broadwell) processors with about 2300 cores in total and a small number of Xeon Phi nodes, based on Knights Landing (KNL). The network fabric is Intel Omni-Path. Storage is GPFS-based at a capacity of 0.5 PB.
- The Cloud Infrastructure for Microbial Bioinformatics (CLIMB) includes a distributed four-site computing system capable of supporting over 1,000 virtual bioinformatics servers. This system consists of hardware with 7,680 cores, 2,304 TB of cross-site replicated storage and 78TB of RAM.
- Warwick Medical School is investing £300K in a 1PB storage server with another 1PB providing backup. Periodic snapshots of live data will be taken and replicated to backup systems. This will be installed end of July 2017. In addition, WMS is improving network infrastructure across the site providing 10 Gb connectivity in some locations with a view to expanding this in the future as and when required.
- The Zeeman Institute SBIDER uses three main computer servers: one based on five Dell Poweredge servers, with 224 CPU cores, 1280 GB RAM & 30TB storage; a file server based on FreeNAS OS and ZFS filesystem which is mirrored for backup, 47TB; and a HP ProLiant DL580 Gen9 server with 1 TB RAM and 64 Intel Xeon E7-8860 2.2GHz CPU cores.

Training

The University of Warwick operates an extensive programme of doctoral training (Masters and PhDs) and CPD to support capacity development across biomedicine and related quantitative disciplines. These include:

- **MathSys** (funded by EPSRC and MRC) trains the next generation of scientists needed to tackle global challenges involving complex, non-linear, uncertain and stochastic systems and includes MSc-level training in fundamental quantitative methods. The majority of each cohort of 10 - 15 have interests in health and medicine, tackling projects as diverse as 'Mathematical modes of the 2013 measles outbreak in Swansea' to 'Bayesian Methods in Health Informatics Data for pancreatic cancer risks' – all PhD projects involve an external partner.
- **OxWaSP** (funded by EPSRC and MRC) trains the brightest graduate statisticians (10 per cohort) in the theory, methods and applications of statistical science for 21st Century data, intensive environments and large scale models. Training includes computational statistics, stochastic simulation, scalable methods and analysis of large complex data, machine learning, time series and stochastic processes. Recent example projects in health-related areas include: 'Particle Monte Carlo methods for clustering multiple data sets' with applications to genomics data, 'Robust Bayesian individualized treatment regimes' (with Novartis).
- **Interdisciplinary Biomedical Research DTP** (funded by MRC and also involving A*STAR in Singapore and the MRC Unit in The Gambia) provides training for 5+ biomedical students per cohort in quantitative and interdisciplinary skills. This DTP focuses on addressing important biomedical questions in the key areas of Quantitative Cellular Dynamics and Molecular Microbiology.
- With its partner NHS Trusts, the University of Warwick also supports Health Education England's integrated clinical academic training by providing related research training and professional development opportunities for clinicians and allied health professionals.

Public Involvement/Engagement

- **UNTRAP** (University/User Teaching and Research Action Partnership) is a partnership between users of health and social care services, carers, and academic staff at the University of Warwick. It promotes and facilitates the engagement and involvement of service users and carers from the local community in research, teaching and governance in health and social care.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr Institute
- MRC Medical Bioinformatics CLIMB – encompasses the following partners: University of Warwick, Cardiff University, Swansea University, University of Birmingham, University of Bath, University of Leicester and The Quadram Institute (a BBSRC research institute)
- Alan Turing Institute founders are the Universities of: Cambridge, Edinburgh, Oxford, UCL, Warwick

NHS and other government organisations

- University Hospitals Coventry and Warwickshire
- Health Education England
- NICE and Joint Committee on Vaccination and Immunisation

Data providers

- CPRD
- NHS Digital
- Public Health England

Industry

- Tech: IBM, Nikon, Jaguar-Land Rover

International

- EU FP7 TRANSFoRm project
- EU 2020 C3-Cloud project
- NIH Human Connectome project

GEOGRAPHIC REGION: ENGLAND – LONDON

- Statements of Interest (Sols) to become a Substantive HDR UK Site were received from the following five ROs within this geographic cluster:
 1. **Imperial College London (ICL)**
 2. **King's College London (KCL)**
 3. **London School of Hygiene and Tropical Medicine (LSHTM)**
 4. **Queen Mary University London (QMUL)**
 5. **University College London (UCL)**
- Due to time/resource constraints, this report contains information only on ROs which expressed interests in becoming HDR UK Substantive Sites. The **Institute of Cancer Research**, which is based in this geographic region, had also submitted an Sols to become a Partner HDR UK but this report doesn't contain any information on its informatics/health data research activities.

Cross-region Strategic Activities

- It's noteworthy that three out of the six **Academic Health Science Centres (AHSC)** in England are based in London (*see also section 2.3.1*), which are nested within wider **Academic Health Science Network (AHSN)** (*see also section 2.4.1*). These are:
 1. **Imperial College AHSC and Imperial College Partners AHSN**
 2. **King's Health Partners AHSC and Health Innovation Network (AHSN South London)**
 3. **UCL Partners AHSC and UCL Partners AHSN**
- The three Centres have come together to discuss plans to establish a **pan-London Trusted Research Environment** (leads include: Professor Tim Hubbard from KCL and Professor Harry Hemingway from the Farr).
- **MedCity** is a collaboration between the Mayor of London, the capital's three AHSC and the Universities of Cambridge and Oxford. Supported by HEFCE and the Mayor of London and launched in 2014 to promote and grow the world-leading life sciences cluster of England's south east, it is promoting life sciences investment, entrepreneurship and industry in the region by:
 1. providing a single front door and concierge service for industry and investors looking for partners, infrastructure and expertise;
 2. facilitating and supporting collaboration across all parts of the sector to turn innovations into commercial products and services;
 3. fostering an environment that support and encourages entrepreneurialism; and
 4. raising awareness globally of the regions rich life sciences ecosystem.

- **DigitalHealth.London** is a collaboration between MedCity, NHS England and London's three AHSNs and supported by the Mayor of London and the AHSCs. This collaboration was formally launched in February 2016 by George Freeman in response to the London Health Commission's 'Better Health for London' report recommendation for the creation of a digital health hub in the city. It has been created to improve quality of care, make healthcare more effective and efficient, and to generate economic growth. The aim is to meet the challenges for industry, clinicians, commissioners and patients by:
 1. creating and supporting networks to build on London's existing critical mass of digital health expertise;
 2. building knowledge between stakeholders to enable greater understanding of challenges, requirements and issues;
 3. identifying digital solutions that meet the needs of patients and clinicians;
 4. building the evidence base of the health outcomes and economic benefits of digital health; and
 5. supporting the collaborative development of an environment that both meets the needs of the health system and allows businesses to develop by tackling issues such as procurement and commissioning.
- For example, the **DigitalHealth.London Accelerator programme** is uniquely positioned to provide in-depth knowledge of the NHS to a cohort of high potential SMEs. The Accelerator aims to speed up the adoption of innovation at scale, ensuring that together greater numbers of patients benefit from emerging new technology, more rapidly. The programme is receiving £1.7M funding from the European Regional Development Fund. The year-long programme will run over three consecutive years to work with around 80 (20-30 in each year) SMEs and support them through meetings and events individually tailored to each business's specific needs, which focuses on:
 1. engagement with clinicians and healthcare experts;
 2. refining products to meet needs;
 3. deepening understanding of the health system;
 4. showcasing in health facilities;
 5. developing business models in order to progress product development; and
 6. access to market.

Imperial College London

Top-level insights

- Imperial College London (ICL) has a number of investments in informatics research (e.g. MRC UK MEDBIO, NIHR HIC, Imperial Data Science Institute) underpinned by excellent multi-disciplinary teams, bringing together complementary skills and expertise in medicine, computer science, data analytics and engineering (with substantial investment from EPSRC).
- ICL has a long track record (e.g. Small Area Health Statistics Unit, Dr Foster Unit, Big Data Analytics Unit) of processing, analysing and drawing knowledge and insights from national health and social care datasets; and patient and population data linked to 'omics and to clinical data (e.g. internationally leading position in the generation, analysis and integration of metabolomics data to other types of data).
- There is a substantial supercomputing resource, with (currently) up to 4 PB of data storage hosted in Imperial's Data Centres, and a commitment to further significant investment over the coming years to continuously develop their HPC hardware.
- Furthermore, the Imperial College AHSC is nested within Imperial College Health Partners (IChP) AHSN; these have aligned strategic visions to support the progression of research from discovery to implementation and dissemination, with an emphasis on training and informatics. This facilitates the pull through of advances in informatics research from the lab bench to the clinical practice (e.g. they aim to develop and deliver a single EHR for NW London).
- Imperial was not involved in the establishment of the Farr Institute but is collaborating with the Farr@Swansea, as part of the MRC UK MEDBIO Medical Bioinformatics award and the Airwave study.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

ICL's major interdisciplinary investments in the areas of biomedical research and population health with relevance to data science include:

- **MRC UK MEDBIO** Medical Bioinformatics award (£6M; *Lead: Elliott; other investigators involved in this award: Butcher, Glen, Nicholson, Guo*) (see section 2.1.2).
- The **Imperial Data Science Institute (DSI)** (*Director: Professor Yike Guo*) is world-leading in the development and application of new software solutions and technologies for data assembly, data analytics and data visualisation for complex, high-dimensional clinical, 'omics and imaging data, and includes development of the tranSMART system.
- The NIHR Imperial BRC is one of the five NIHR BRCs involved in the **NIHR HIC programme** (see section 2.2.2), leading on Acute Coronary Syndrome (*Dr Jamil Mayet based at ICL is the lead investigator for this theme*).

Knowledge/Expertise

- ICL is the only UK University to focus exclusively on medicine, science and engineering, as well as business.
- It has strong inter-disciplinary research and training across biomedicine, computing, engineering, mathematics and statistics underpinned by substantial computing resources and wide-range of relevant expertise and critical mass.
- In addition to the bioinformatics, data analysis and visualisation expertise, Imperial has long track record of processing, analysing and drawing knowledge and insights from national health and social care datasets; and patient and population data linked to 'omics data and to clinical data.

Compute

- Imperial offers a secure and productive environment for analysis in a flexible hybrid computational infrastructure using a range of HPC and networked servers (Imperial HPC, UK MED BIO, DSI) with (currently) up to 4 PB of data storage hosted in its Data Centres. A more detailed description of their computational resources can be found under the 'e-Infrastructure' section.

Data

- ICL's **MRC-PHE Centre for Environment and Health** (£6M; *Elliott*) includes the UK Small Area Health Statistics Unit (SAHSU), which holds over 300 M geocoded records including national data on Hospital Episode Statistics (HES), mortality, cancer registrations and birth records.
- It has established extensive highly phenotyped patient and population cohorts that include:
 - The **London Life Sciences Prospective Population Study** (“**LOLIPOP**”) – a major UK research study investigating the mechanisms underlying heart disease stroke, diabetes, obesity and other major medical problems. The study has collected lifestyle information and samples from over 30,000 people in North West London;
 - The **Airwave Health Monitoring Study** has recruited more than 52,000 participants, hosted in the Farr@Swansea secure servers for inclusion in Dementia Platform UK (DPUK).
- The **NIHR HIC Acute Coronary Symptoms dataset** contains data from the NIHR HIC partners and is stored at Imperial College Healthcare NHS Trust's research warehouse. It is an anonymised dataset derived from routinely collected NHS patient data (250K patients). These are secondary care- local deep episodic patient-level data including: A&E records; pathology tests; drug treatment; imaging (structural and functional); revascularisation (percutaneous, surgical); echo imaging; transfusion.

Translational and innovation capabilities

Industry links

- ICL's **Confidence in Concept initiative** encourages pull-through of new diagnostics, devices and interventions, including support for development through Imperial Innovations, and through industrial partnership facilitating external funding. In 2014/15, Imperial won the highest volume of industry research funding of all UK universities (£46M).
- Industrial partners through the MRC UK MEDBIO Medical Bioinformatics award include: Waters Corp., Bruker Biospin, Huawei Technologies Co. Ltd., Thomson Reuter, Astra Zeneca.

NHS links

NIHR Biomedical Research Centre

- The **NIHR Imperial BRC** (£90M; *Director: Jonathan Weber*) is a partnership between Imperial College and the Imperial College Healthcare NHS Trust. This BRC has a cross-cutting theme in Informatics and Biobanking, which builds on major strengths in informatics, analytics and data linkage, patient and population cohorts and tissue biobanking, with aim of providing the resources and infrastructure to carry out experimental medicine studies, prognostic research and biomarker discovery, validation and qualification across the BRC. Furthermore, Imperial College in collaboration with the NHS Trust is setting up a new and exciting project – the **Universal Consent project** – across the Trust to obtain consent for research, linkage of health data and biological samples from patients attending the Trust. These data will feed into the Imperial BioResource. The project roll out across the Trust was planned for 2017 and will include ISO27001 Information Security Management Accreditation and ongoing compliance at IG standards, level 3.

It also builds on existing BRC-BRC informatics-omics collaborations nationally, including the **NIHR HIC** and the **NIHR Bioresource**. The Imperial BRC has been leading on the acute coronary syndrome theme (*Mayet*) responsible for the collection, storage and analysis of this large dataset (>250K patients). Together with researchers from Imperial's Data Science Institute, they have been using mathematical modelling for prediction of arrhythmia or vulnerable atherosclerotic lesions, and mapping generic abnormalities to heart morphology/function.

Academic Health Sciences Centre

- The **Imperial College Academic Health Sciences Centre (AHSC)** (*Executive Director: Professor Jonathan Weber*) is a partnership between Imperial College and three NHS Trusts:

1. Imperial College Healthcare NHS Trust
2. Royal Brompton & Harefield NHS Foundation Trust
3. Royal Marsden NHS Foundation Trust

which is creating a critical mass of 21,500 clinicians and other NHS staff, researchers and academics to drive innovation and improved care for the direct benefit of over 1.1 M patients each year in West London.

The AHSC **themes are aligned with domains that reflect the translational pipeline of the NIHR Imperial BRC.**

Academic Health Sciences Network

- The Imperial College AHSC is nested within **Imperial College Health Partners (IChP) Academic Health Science Network (AHSN)**. The strategic visions of both the Imperial College AHSC and AHSN have been aligned to support the progression of research from discovery to implementation and dissemination, with an emphasis on training and informatics.
- The Imperial AHSC, with Med City and the other AHSCs and AHSNs in London are also part of the **DigitalHealth.London initiative**, launched in Feb 2016.

Collaboration for Leadership in Applied Health Research and Care

- **NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North West London** and the Department of Computing at Imperial have designed a bespoke web application, the Web Improvement Support for Healthcare (WISH). This is an online framework that enables data capture and live reporting for quality improvement projects. The platform has been used by over 70 improvement projects in London hospitals over the past seven years.

NHS 100,00 genomes project | Genomics England Ltd

- The ICHT is the lead delivery partner for the **West London NHS Genomic Medicine Centre**, working in partnership with the Royal Marsden, Royal Brompton and Chelsea & Westminster NHS Trusts, and the AHSN. The project is supported by the AHSC informatics platforms.

Opportunities

Trusted Research Environment

- Imperial College and Imperial College Healthcare NHS Trust have developed a **joint secure infrastructure/environment** for analysis and storage of sensitive health data for research (established through the NIHR HIC) with appropriate anonymisation/pseudonymisation and IG processes to provide ready access to bona fide researchers.

The data are legally stored and used in a federated secure environment, allowing linkage between data captured routinely in the NHS, NIHR HIC data, Dr Foster data and other clinical and research datasets, to provide a richer, precision medicine focussed resource. The platform is planned to receive NHS Digital IG Toolkit Level 3 during 2017.

Building on the NIHR HIC project, Imperial plans to develop a research engine to extract data from Cerner and other databases and link to the patient record. They also had discussions with Google DeepMind about potential to work collaboratively toward this.

- This infrastructure is being developed in parallel to a Trust-led initiative to develop a **single EHR for NW London**. In aligning these initiatives, the AHSC aims to create a uniquely comprehensive, rich population data set for research, housed in a secure, managed-access infrastructure.

Consent via NHS records

- There are ca. 1 M patient contacts per year at the Imperial College Healthcare NHS Trust, but currently the research potential of this massive clinical resource is largely untapped. Through, the NIHR Imperial BRC, they will be rolling out the **DISCOVER programme (Elliott)**, which seeks consent for research follow-up via NHS records; and linkage to biological samples for patients attending the Imperial College Healthcare NHS Trust (over 1 M patient visits p.a.). This will open the avenue to more phenotypic-rich data to be used for research. This initiative complements the NIHR BioResource of which Imperial is a partner.
- Furthermore, in collaboration with Oxford AHSC, a common universal consent process is also being piloted across both AHSCs.

Potential gaps

- From the body of evidence collected so far for the purposes of this report, it seems that Imperial College has just started to expand its research on citizen-driven/digital health, while they seem to have the resource and expertise to delve even deeper into learning health systems research.

Other relevant capabilities

Research themes

Imperial has substantial investments from basic science to point-of-care that include:

Discovery science

- The **MRC London Institute of Medical Science** is performing pioneering work on fundamental epigenetic mechanisms of disease.

Precision medicine

- The **MRC-NIHR National Phenome Centre** (£10M; *Nicholson, Elliott, Holmes*) is at the forefront of developing high-throughput methods for metabolic phenotyping (NMR and mass spectrometry) of large patient and epidemiological cohorts. It is also linked to the **Clinical Phenome Centre** (a core facility of the Imperial College BRC) for experimental/stratified medicine research.
- **EPSRC Centre for Mathematics of Precision Healthcare** (£2M; *Barahona, Guo, Rueckert, Darzi*) brings together researchers from across Imperial to develop and apply next generation tools, encompassing graph theory, dynamical systems, stochastic processes, statistical learning and optimization, to extract meaning from complex medical data.
- The **STELAR** (Study Team for Early Life Asthma Research) consortium (*Custovic*) aims to identify novel endo-types of childhood asthma using machine learning approaches.

Population and Public health

- **MRC-PHE Centre for Environment and Health** (*Elliott*) was formed in June 2009 as a partnership between Imperial College (lead institution) and King's College London, funded by the MRC and Health Protection Agency (now PHE). The Centre incorporates the **UK Small Area Health Statistics Unit (SAHSU)** at Imperial College, the Environmental Research Group (ERG) at King's College London, in addition to bringing together leading researchers from four (of six) departments of the Faculty of Medicine at Imperial College, and researchers at King's College, St George's University of London and London School of Hygiene and Tropical Medicine. The Centre forms a multi-disciplinary research cluster, covering a wide range of techniques and approaches; using advanced geographical information systems and statistical modelling techniques, combined with experimental data, biomarker and mechanistic studies, and analyses of large population cohorts to explore environmental health problems of key public health and scientific importance.
- The **UK Small Area Health Statistics Unit (SAHSU)** at Imperial College is a nationally and internationally recognised institution for research into environment and health. SAHSU's work includes substantive epidemiological enquiries of environmental health problems and methodological research. SAHSU studies may involve (as necessary) environmental modelling and monitoring as well as biomarkers (of exposure, effects and genetic susceptibility) in addition to the analysis of routine health statistics, to aid interpretation of causal inference. The Unit now forms a core part of the MRC-PHE Centre for Environment and Health held jointly between Imperial College London and King's College London set up in 2010.
- The **Dr Foster Unit** (£1.5M; *Aylin*) was established in 2002, as a partnership with Dr Foster, and it sits within the Department of Primary Care and Public Health in the School of Public Health at Imperial College. Its research is funded through a variety of grants, with a key source of funding coming from Dr Foster – a wholly owned subsidiary of Telstra. The Unit also receives NIHR funding for its work with the Patient Safety Translational Research Centre (PSTRC) and the Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR HPRU HCAI and AMR) and to investigate readmissions in heart failure and chronic lung disease. NIHR has also funded projects involving the mortality alerts the Unit sends out each month and obstetric care.

Citizen-driven/Digital health

- The **Adaptive, Real-time, Intelligent System to Enhance Self-care (ARISES)** research project (£1.3M; EPSRC) aims a system that will run on a smart phone locally and collect data from multiple sources to deliver an intervention to the patient that allows self-management of chronic disease.
- The **Novel Wearable Technology for Early Detection of Exacerbations in COPD** research project (£0.9M; EPSRC) aims to create a novel wearable wireless technology- compose of a sensing unit of approximately the size of a pound coin to be worn by the patient in the neck, and a mobile phone application- that will be able to monitor COPD patients continuously, and automatically provide early detection of potential exacerbations, in order to inform patients and/or their doctors, so that these can be treated promptly to minimise their likelihood of progression to higher levels of severity.

Disease areas

- Imperial's research focuses on several disease areas including: cancer, cardiovascular, diabetes, infectious diseases and antimicrobial resistance.

Analytics

In addition to the Imperial Data Science Institute and the EPSRC Centre for Mathematics of Precision Medicine, other investments and activities in health data analytics at Imperial College include:

- The **Dr Foster Unit** has expertise in managing and analysing 2.8 billion health records, including: NHS hospital, GP consultations, cancer registry and international health data. Their work forms the basis of the near **Real-Time Monitoring System**, currently used by 70% of English NHS Acute Trusts to monitor case-mix adjusted health outcomes at individual patient level.
- The **Big Data Analytics Unit** in the Centre for Health Policy (*Darzi*) provides high powered analytic capability for data integration, linkage and analysis, including of sensitive NHS data.
- The **MRC Centre for Outbreak Analysis and Modelling** (£2.7M) (*Ferguson, Ghani*) analyses worldwide data on emerging or newly resurgent infectious disease threats (e.g. SARS, H5N1 avian influenza, Ebola) and endemic infections of major global health significance (e.g. malaria, TB, HIV).
- The **EPSRC Smartheart programme** (£5.1M; *Rueckert, Guo*) is applying artificial intelligence and machine learning for computer-aided detection and diagnosis of fetal abnormalities and cardiovascular disease.

e-Infrastructure

Compute

- There is a substantial supercomputing resource, with (currently) up to 4PB of data storage hosted in Imperial's Data Centres, and a commitment to further significant investment over the coming years to continuously develop their HPC hardware. All HPC hardware is in secure, limited-access purpose-built Data Centres on the South Kensington campus and a second site (JISC Datacentre, Slough), managed centrally.
- The ICT HPC group have over £5M investment in HPC via three main HPC systems. These are:
 1. Cx1 (general purpose computer cluster, covers a significant amount of the computing requirement with a large and varied workload) – a large “Beowulf” PC cluster system, with gigabit ethernet connectivity, and a range of different node types (including a number of GPU nodes) using common system images and user environments. Currently cx1 has 1395 nodes and 13,558 cores including the dedicated UK MED-BIO nodes. The large size of the system means that it can be very responsive to complex individual user demands providing additional resources to those of UK MED-BIO.

2. ax4 – is a very large shared memory system (SGI UV2000), specifically for genomics. It has 15 TB of RAM and 1280 cores, plus 1.5 PB of fast local RAID disk. The group also offers access to a large scale, massively parallel processing (MPI) system.
 3. Data Science Institute (DSI) has a dedicated Data Centre oriented toward data exploration, analysis and visualisation. Built around a fully redundant 40Gb/s network providing 18Tb/s of switching capacity, with private redundant clouds based on Open Stack powered by >50 high density Xeon-based computation servers. Database clusters are maintained (MongoDB, Maria and PostgegreSQL) with distributed computation clusters based on Spark, and a centralised storage holding 600TB of mirrored space comprising >12TB of processing RAM. DSI provides GPU based K80 clusters. Other activity includes fully secured Qualys A+ webhosting and software stack training. Equipment is located on the South Kensington campus. The KPMG Data Observatory houses a 313 degree surround visualisation system allowing exploration of data and running of scenarios in real time.
- Dedicated biomedical HPC and storage was put in place through the MRC UK MED-BIO. This includes a flexible hybrid computational infrastructure with a >3,040 core cluster and cache-coherent memory SGI UV system for memory-intensive work (680 cores, 8.5 TB RAM, each with 0.5PB dedicated local scratch); project-specific environments run on 8 high memory servers. It has a multi-petabyte high performance tiered storage system/Infinband/10GB backbone, supporting 500 TB GPFS, 2 PB WOS object storage and 2PB tape archiving (on two sites – load balancing and disaster recovery).

Platforms

- Imperial (*Guo*) is leading the development of the **tranSMART platform**, an open-source, community-driven knowledge management platform for translational medicine that can enable pre-competitive and private data sharing and foster collaboration. It also provides tools for: loading a variety of high-content data types, chemical, biological and 'omics research and clinical data, from a range of sources; investigating the relationships between genetic and phenotypic data for cohorts of patients; and assessing analytic results in the context of published literature and internal work.

The platform is managed by the tranSMART Foundation (*Guo*, CTO), a non-profit organization. It is a project that is being collaboratively developed by more than 100 computer and clinical scientists from more than 20 organisations from around the world.

- The Imperial Data Science Institute (*Guo*) is also leading the development **eTRIKS** (€ 23M ; 2012 -2017) informatics/KM platform, to support translational research projects funded through the Innovative Medicines Initiative (IMI), Europe's largest public-private initiative.

Training

- Imperial has a large and well established training programme as well as mentorship and career development support for pre- and post-doctoral students and faculty.
- The training programmes include: **MSc or MRes in Computing** (Machine Learning), **Bioinformatics & Theoretical Systems Biology, Statistics, Genomic Medicine**, Epidemiology; and are starting a new **MSc in Data & Health Science** (from October 2017).

Public Involvement/Engagement

- The **Imperial College Patient Experience Research Centre (PERC)** research programme is being implemented into clinical practice across the AHSC to improve the collection and analysis of patient experience data and to support service improvement and redesign; ICHT has developed a PPI strategy with PERC.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- NIHR Imperial BRC
- MRC Medical Bioinformatics award (UK MED BIO) partners: Imperial College – host institution; Institute of Cancer Research, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI); Farr Institute (Farr@Swansea (CIPHER)); MRC Clinical Sciences Centre; MRC Human Nutrition Research (MRC-HNR).
- UK Biobank Cardiovascular Consortium
- MRC-NIHR National Phenome Centre
- King's College London – partnership in SASU

DH & NHS

- NIHR BRC HIC partners: Oxford, Cambridge, King's College, UCLH and Imperial College
- NIHR School for Public Health Research
- NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Northwest London
- Genomics England – West London Genomic Medicine Centre
- Imperial AHSC and AHSN (Imperial College Health Partners (IChP))
- DigitalHealthLondon
- Oxford AHSC
- NIHR CLARHC North West London
- North West London Local Clinical Research Network (LCRN)
- School of Public Health at Imperial College London

Data providers

- PHE
- NHS Digital
- ONS

Industry

- Waters Corp.
- Bruker Biospin
- Huawei Technologies Co. Ltd.
- Thomson Reuter
- AstraZeneca

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:

International collaborations

- tranSMART foundation
- eTRIKS consortium
- EU PhenoMeNal on standards in metabolomics with EMBL-EBI
- EU FP6 GABRIEL
- STELAR consortium
- INTERMAP: International Population Study on Macronutrients and BP
- COSMOS (COhort Study of MObile phone uSe and health)
- Human Connectome Project (ERC Synergy Grant) and
- International Network of Phenome Centres (Nicholson)
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR

King's College London

Top-level insights

- King's College London (KCL) has strengths on processing clinical data from multiple sources, including EHRs, large medical imaging datasets and mobile devices; and integrating these to their BioResource (multi-omics) and external non-health databases. In addition, they have experience in developing and operating platforms to support research such data.
- KCL's is working very closely with its partner NHS Trusts and its key health and biomedical informatics research activity is being delivered jointly mainly through NIHR support. For example, the Clinical Record Interactive Search (NIHR CRIS) system is the UK's largest and most extensive electronic clinical informatics database for mental health, was developed at the NIHR Maudsley BRC.
- Furthermore, KCL's partner NHS Trusts have over 4 M patient contacts/year, with EHR data suitable for informatics approaches. Their ability to provide leadership in clinical informatics – including health records, genomics and medical imaging – is supported by the two NIHR BRCs. The Maudsley BRC has produced ground-breaking research using EHR and mhealth, while the Guy's & St Thomas' BRC has world-leading medical imaging informatics and excellent genomics, well-connected to national efforts such as the NHS 100k Genomes Project delivered by Genomics England.
- Also through the Guy's & St Thomas' BRC, they are involved in the NIHR HIC programme.
- KCL's research is supported by a new £3.6M investment in an HPC system (Rosalind); and they also have access to a trusted research environment that allows secure data linkage.
- KCL is one of the few UK research organisation which has substantial activity and is leading international efforts in mobile/digital health (e.g. RADAR-CNS programme) to deliver clinically useful data from smartphones and wearables.
- As one of the UK's leading AHSC, King's Health Partners (KHP) AHSC aims to become an international leader in the field of optimised clinical and research informatics. The AHSC is nested within an AHSN: Health Innovation Network (South London).
- KCL was not involved in the establishment on the Farr Institute but has developed connections with the Farr mainly through their collaboration on the MRC eMedLab Medical Bioinformatics award. They also make significant contributions to the ATI (e.g. contributing to an ATI report on big data in imaging; and an ATI-Intel programme on mHealth). KCL is a core member of the Crick Institute.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- KCL's key biomedical informatics research is being delivered jointly with their NHS Trust partners mainly through major NIHR BRC investments (*see also section further down on 'NHS links'*). These are:
 - The NIHR Maudsley BRC's **Centre for Translational Informatics** has been established to create connections between the NHS IT and academic health informatics and acts as a catalyst for the industry and academic partnerships, delivering a clear pathway for informatics-based interventions into routine practice. For example, the Centre aims to drive digital innovation in mental health, finding new ways of improving healthcare using digital technologies and accelerating their delivery to patients. It is led jointly by South London and Maudsley NHS Foundation Trust and the Institute of Psychiatry, Psychology & Neuroscience at King's College London.
 - The **Clinical Record Interactive Search (CRIS)** (*Lead: Professor Robert Stuart*) is the UK's largest and most extensive electronic health record data resource developed with support by the NIHR Maudsley BRC (currently >270k records, >70 publications). The CRIS system allows rich and pseudonymised clinical data held in NHS Trusts to be used for research.

The CRIS system has been deployed in (mental health) NHS Trusts since 2014 under the **Dementia (D)-CRIS** (collaboration programme) to significantly extend this research capability. The participating NHS Trusts include: Camden and Islington, Oxford Health, Cambridgeshire and Peterborough, West London Mental Health and South London and Maudsley.

- The NIHR GSTT BRC is one of the five **NIHR HIC** partners and is leading on the renal transplantation research theme and the (de-identified) data collection (for this theme) from across four HIC participating centres.
- KCL is one of the key partners of the MRC-funded Medical Bioinformatics eMedLab award (*see also section 2.1.2*).

Knowledge/Expertise

- KCL has very strong track record and expertise in developing and operating platforms and software solutions to support research using EHR data and data from mobile devices.
- This is in addition to more traditional robust expertise in genomics and imaging data collection and analysis; and their linkage to data from other sources.

Compute

- King's and Maudsley and GSTT Charities have jointly invested in **HPC** (£3.6M put in place on February 2017). This facility is co-located with other HPC investments (MRC eMedLab and the Maudsley's clinical systems) in the JISC Shared Data Centre in Slough.

Data

- The **D-CRIS system**, mentioned previously, enables large datasets from the participating NHS Trusts to be pooled so that research can be conducted at scale, providing researchers with access to 1M pseudonymised patient records.

- Through the **NIHR HIC**, KCL has access to renal transplantation data from across four HIC participating centres. This includes secondary care data from patients (currently estimated 6,500 unique patients) who have undergone kidney, or pancreas and kidney, transplantation; as well as information on recipient donor demographics, medical history, transplant details, pathology/biopsy and HLA results, medication and follow up.
- The King's Health Partner's **i3 Diabetes programme** has been working with industry partner Aridhia to join up primary and secondary care data for diabetes to help drive improvements in research and ultimately care pathways for patients. 93 GP practices are in the process of agreeing data sharing across Lambeth and Southwark localities.
- KCL has access to the **Lambeth DataNet**, which is an anonymised primary care database for 380,000 registered Lambeth patients. In partnership with Queen Mary University of London (QMUL), they have combined this dataset with the East London primary care data to create a resource of 1.1M.
- KCL hosts a substantial number of cohorts including local and national disease registers; and cohorts focused on life-course and education; genetic analysis and imaging-genomic analysis (e.g. the deeply characterised Twins UK).

Translational and innovation capabilities

Industry links

- KCL has partnerships with digital technology and pharma companies. Its current portfolio of **tech industry partnerships** includes: Intel, SoftwareAG, Ixico, the tranSMART foundation, FitBit, MacLaren Applied Technologies, The Hyve, Vibrent Health, Findwise, Precognox and Elastic.co, FACE Ltd, Datalytics, CARE-IS Ltd.
- They have collaborations with DXS-UK Ltd in the implementation of clinical decision support tools into GP systems for trial intervention; and have links with EMIS, TTP and InPS.
- Strategic partnerships in imaging sciences for 'smart' scanning include Siemens GE and Phillips Healthcare, with scientists from these companies embedded in King's research groups.
- They have recently signed a partnership framework agreement with Unilever which will provide them with substantial consumer data.
- KCL is leading on **EU Innovative Medicines Initiative programmes** (e.g. RADAR-CNS and EMIF).

NHS links

NIHR Biomedical Research Centres

- KCL is a partner in two **NIHR BRCs**:
1. The **NIHR Maudsley BRC** (£66M; *Director: Professor Matthew Hotopf*) is a partnership between the Institute of Psychiatry, Psychology & Neuroscience at King's College London and the South London and Maudsley NHS Foundation Trust.

The Maudsley BRC has invested heavily in informatics infrastructure to support:

- a. **Bioinformatics and statistics** (*theme leads: Professor Andrew Pickles and Dr Richard Dobson*): bioinformatics and statistical translational research; and to develop biomedical infrastructure complementary to the overall strategy of the BRC. This is done through the integration of rich clinical data from patient records with large variable datasets including genomics, epigenetics, proteomics, neuroimaging and others, to further describe psychiatric disorders; and

b. **Clinical and population informatics** (*theme lead: Professor Robert Stewart*)

This builds on two prominent achievements of the BRC:

- the CRIS system (mentioned above) and
- the SouthEast London Community Health Study – the UK’s largest and most comprehensive study of urban mental health

Activities under this theme aim to extend clinical and population mental health data resources through online recruitment platforms and enhanced clinical databases and then, apply these resources for improving physical health outcomes, supporting precision psychiatry and novel therapeutics, and delivering informatics based interventions. They also provide an appropriate platform to export data generation/processing tools through a national e-network for mental health informatics.

2. The **NIHR Guy’s and St Thomas’ (GSTT) and KCL BRC** (£64.4M; *Director: Professor Graham Lord*) is a partnership between KCL and Guy’s and St. Thomas’ NHS Foundation Trust.

This BRC has a **Research Platforms theme** (*theme lead: Professor Mark Peakman*), which includes: Immune Monitoring, Genomics, Translational Bioinformatics, Advanced Therapeutics Manufacturing (GMP) Unit, BioResource, NIHR HIC, Statistics/Biostatistics and the BRC Clinical Trial Hub.

It also has a **Genomic Medicine theme** (*theme lead: Professor Richard Trembath*), which utilises patient cohort datasets, longitudinal population studies (e.g. Genomics England; UK BioBank; TwinsUK; East London Genes and Health); and infrastructure within the NIHR, including: the Rare Diseases Translational Research Collaboration (*Simpson*); the BioResource (Spector); the NIHR HIC (*Hubbard/Lord*); and the BRC Genomics and Bioinformatics Platforms.

Theme leverage includes: the MRC Clinical Research Infrastructure award (£2M) for Single Cell Functional Proteomics and Genomics; the MRC eMedLab partnership (£8.9M); and the HEFCE-Catalyst Centre for Population Genomics (£5.1M partnership) theme.

This BRC provides underpinning for the **NIHR HIC** leading the transplantation research theme (*Chowdhry*) and developed a Data Mart in TransMart.

Academic Health Science Centre

• The **King’s Health Partners (KHP) Academic Health Science Centre (AHSC)** (*Director: Professor Sir Robert Lechler*) is a partnership between King’s College London and three NHS Trusts:

1. Guy’s and St Thomas’ NHS Foundation Trust
2. King’s College Hospital NHS Foundation Trust
3. South London and Maudsley NHS Foundation Trust

Building on the success of the KHP Online, a web portal which allows joining up EPR right across the three partner Trusts and making this information available to clinicians across King’s Health Partners, they have further developed it to link to GP practices in Southwark and Lambeth through the **Local Care Record**. This is the first patient information sharing system in the UK to receive an official accreditation from medical information system supplier EMIS Health.

KHP Online has led to immediate improvements in the management and care of patients and service users. It was also featured as a case study by Dame Fiona Caldicott in her Information Governance report to the Secretary of State for Health in December 2014.

Academic Health Science Network

- As one of the UK's leading AHSC, KHP aims to become an international leader in the field of optimised clinical and research informatics and have established the KHP Informatics Committee, which oversees and coordinates informatics activities across partners. The AHSC is nested within an AHSN: Health Innovation Network (South London).

Collaboration for Leadership in Applied Health Research and Care

- Novel decision support systems are being developed for stroke patients and their clinicians within the **NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) South London** with Care IS Ltd using the WISH (in collaboration with Imperial's NIHR CLAHRC Northwest London).

NHS 100,000 Genomes Project | Genomics England

- The NIHR Maudsley BRC has developed a CRIS-like capability for the use of general hospital Trusts, which pulls in data from multiple sources, de-identifies it, and applies text mining apps to extract meaning. This processing pipeline forms the basis of their commitment to Genomics England's clinical phenotyping pipeline, which they lead (*Professor Tim Hubbard is the Head of Genomics England Genome Analysis*), and has now been successfully installed in collaboration with the UCLH and within King's College Hospital.
- KHP AHSC hosts the **South London Genomics Medicine Centre**. In addition, KHP researchers lead and co-lead multiple GeCIPs, including the analytical and EHR-based domains.

Opportunities

- KCL is building on their strengths to create some KHP-wide broad, deep and consistent datasets to support research and clinical care.
- They are leading on the efforts (*Hubbard*) to establish a London-wide Trusted Research Environment to be hosted at Farr@London.

Potential gaps

- KCL does not seem to have any significant public involvement/engagement activity related to health data science.

Other relevant capabilities

Research

Discovery Science

- KCL has substantial in house data generation capacity, in sequencing and around functional genomics of cellular systems, including an **MRC Single Cell Functional Genomics** facility.
- They also work in partnership with the Sanger on the HipSci consortium for high-throughput cellular manipulation.
- Wellcome Trust/EPSRC **Centre for Medical Imaging** (CMI, £12M), which has been funded to break down the barriers between biology, medicine, engineering, physics, mathematics, computer science and chemistry

Precision Medicine

- KCL researchers lead efforts in the following **MRC Stratified Medicine Consortia**:
 - **AIM-HY** (£3.4M; *Chowienczyk*): ancestry and biological Informative markers for stratification of hypertension;
 - **RA MAP** (£4.5M; *Cope*): treatment stratification of early rheumatoid arthritis; and identification of immunological changes that occur as early as the disease develops into a more chronic one;
 - **STRATA** (£5M; *Kapur*): Schizophrenia Treatment Resistance and Therapeutic Advances, combines neuroimaging and genetics studies to investigate clinical outcome in psychosis;
- In addition, KCL researchers are **co-investigators** in the following **MRC Stratified Medicine Consortia**: **PSORT** – psoriasis; **COPD-MAP**; and **MASTERMIND** – type2 diabetes.
- Other activities that KCL leads in this area include:
 - **SmartHeart** (EPSRC £2.1M): next-generation cardiovascular healthcare via integrated image acquisition, reconstruction, analysis and learning;
 - **EU-AIMS** (IMI €36M, led by Roche and KCL): longitudinal autism research cohorts to identify biomarkers for stratification and surrogate endpoints indicative of change;
 - **dHCP** (ERC €15M): the Developing Human Connectome Project, linking advanced imaging (DTI/tractography, resting state networks), clinical and behavioural data;
 - **VP2HF** (€1.2M): computer model derived indices for optimal patient-specific treatment selection and planning in heart failure.

Citizen-driven health

- The **RADAR-CNS** (EU IMI2 €25M) programme aims to derive clinically useful data from smartphones and wearables and develop an open source mHealth platform for monitoring devices that provides core capability for other programmes.
- The **TOHETI** (Transforming Outcomes and Health Economics Through Imaging) (KHP £10M) programme looks cohesively at a patients' journey across different care pathways and includes a variety of stakeholders including patients to co-design and implement changes which will deliver the greatest patient benefit
- The **CONSULT** (EPSRC £1.4M) programme will combine wireless 'wellness' sensors with intelligent software running on mobile devices, to support patient decision- making, and thus actively engage patients in managing their healthcare.;
- The **Kings@Home** is developing video-EEG telemetry for home-based diagnostics of neurological disorders and won the 2013 NHS Innovations Challenge Prize. It is now a clinical service at KCH;
- The **IMPARTS** (Integrating Mental & Physical Healthcare) programme has created an informatics system to facilitate routine collection of patient-reported outcomes, with real-time feedback to guide clinical care.
- The **Weston Programme for Family-Led Research** (Garfield Weston Foundation, £1M) programme aims to allow parents of sick children to define and carry out research, e.g. led to recently published study of environmental influences on new mothers using geo- locating smart-phone apps and walking diaries.

Public and Population Health

- The **Stroke and Air Pollution** (STAR) informatics aims to link stroke, air pollution and primary care data for research.

- The **eLIXIR** (MRC partnership grant £1m) is a project linking routine clinical records and patient samples to improve UK research capability in following life course of physical and mental health disorders, initially deployed across KHP Hospitals before national expansion.

Learning health systems

- **TRANSFoRM** is a 21 partner FP7 project led by KCL, which has developed a LHS infrastructure for phenotype-genotype studies, efficient trials and diagnostic decision support.
- **H2020 KConnect** is developing semantic search technologies for medical information.
- The **Children and Young People's Health Partnership** (GSTT, £6m) is developing a paediatric care system linking hospital and primary care in South London for collecting and analysing data on process and health care implementation.

Disease areas

- King's College's research focuses on several disease areas and in especially on: mental health, immunology and cardiovascular disease.

Analytics

- KCL investigators have developed state of the art tools for ingestion of NHS structured and unstructured data, which include optional de-identification, **natural language processing** (NLP) and semantic annotation; as well as methods to model treatment response and disease trajectories from EHR and population data; and use EHR data for efficient trials.
- They are also developing **AI approaches applied to medical imaging**, such as new algorithms for image analysis, multi-modal image fusion, imaging genetics and motion correction.
- **CogStack**: is an open-source information retrieval toolkit developed at Maudsley BRC and KCH to facilitate searches, NLP (using Elasticsearch), analytics and visualisation of EHR data. CogStack has been deployed at KCH to provide clinical alerting and query access to free text and structured information from its EHR (9m documents, ~250m structured data items and ~20m patient transactions) in near real-time. De-identification functionality allows the platform to be used for research as well as in the clinic. CogStack is deployed outside KHP at UCLH and being extended to SLaM.
- KHP have developed an award winning **Local Care Record platform**, which links primary and secondary care systems for clinical care and is now being extended into social and community care and across the South East London providing a platform to deploy decision support applications.

e-Infrastructure

Compute

- The **Rosalind HPC infrastructure** has a total of 3,912 compute cores connected to a 600 TB lustre high performance file system and a 2 PB Ceph distributed object store via infiniband and 10GbE networking. The compute cluster has a mixture of high and low memory nodes ranging from 2Gb-20Gb/core (average 11Gb/core). 3,016 cores operate as a bare-metal HPC cluster using the open grid scheduler. The remaining 896 cores operate as an OpenStack private cloud allowing provision of projects specific environments through virtualisation.
- KCL has access to the MRC eMedLab private cloud compute and data storage cluster as a consortium partner. The compute cluster has 6,000 cores and 4.3PB bulk storage that can be configured as bespoke virtual servers to process large bioinformatics datasets.

Trusted Research Environments

- The **Safe Haven Clinical Data Linkage Service (CDLS)** provides a trusted environment to link EHR records with other clinical data sources to generate integrated views of patient journeys for research analysis. Multiple linked datasets that have been generated include these between Lambeth DataNet and CRIS; and between CPRD and HES.

Data Platforms

- The **Clinical Record Interactive Search (CRIS) system** is the UK's largest and most extensive electronic clinical informatics database for mental health (also mentioned previously in this section). Applications to access CRIS and the analyses carried out using CRIS are closely reviewed, monitored and audited by a CRIS Oversight Committee, which carries representation from the SLaM Cauldicott Guardian and is chaired by a service user. The CRIS Oversight Committee is responsible for ensuring all research applications comply with ethical and legal guidelines.
- As part of the **NIHR-HIC** partnership, GSTT has developed a **data mart** platform in tranSMART which contains data from the five participating BRCs for the 'transplantation' theme that it leads.
- The **Lambeth DataNet** is an anonymised primary care database for 380,000 registered Lambeth patients hosted at GSTT. A toolkit is being developed to allow it to be integrated with distributed heterogeneous data sources for phenotype-genotype studies, real-time safety monitoring, prescribing, cancer screening and service use. This includes effective methods for trial recruitment, novel trial designs, diagnostic decision support (building on participation in EU FP7 TRANSFoRM project), and Patient-reported Outcome Measures (PROMs) and remote monitoring.
- The KHP **Diabetes and Obesity Data Warehouse** was created within CDLS integrating data from GSTT, KCH, 15 GPs in South East London and elsewhere as part of the KHP Diabetes Modernisation Initiative (DMI). The warehouse design allows testing of data analytics to drive clinical transformation, performance and better co-ordination of care. CDLS will also be used to integrate physical and mental health EHR data to create a KHP-wide implementation of CogStack.
- **Healthlocker** is a patient health record platform which provides: patient access to components of their EHR (e.g. careplans, correspondence, medication, appointments); an interface for patient reported outcomes via smartphone apps, wearable devices; delivery of information and e-health interventions.

Training

- KCL hosts several training and capacity building programmes which will be potentially well aligned with the HDR UK including an **MSc in Data Science**.
- Their **MRC Doctoral Training Partnership** includes core training in informatics and quantitative approaches, and offers studentships in areas such as genomics, computational imaging and health records research.
- Their BRC-supported Bioinformatics Platform provides training for all staff on genomics and bioinformatics, programming and data analysis, in addition to providing expert support for individual projects.

Public Involvement/Engagement

- KCL does not seem to have significant public involvement/engagement activity related to health data science.

Partnerships

- A number of partnerships and collaboration in national and international consortia are mentioned in detail in the previous sections.

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- MRC eMedLab – partnership between Crick, KCL, LSHTM, QMUL, UCL, LSHTM, EMBL-EBI and Sanger
- Alan Turing Institute
- Imperial College London

DH and NHS

- NIHR BRCs – Maudsley BRC and Guy's and St Thomas's BRC
- NIHR HIC
- NIHR CLAHRC South London
- Genomics England – South London Genomics Medicine Centre

Data Providers

- Public Health England
- EMIS, TTP and InPS

Industry

- Pharma: Unilever
- Tech: Siemens GE, Phillips, Intel, SoftwareAG, ixico, FitBit, MacLaren Applied Technologies DXS-UK Ltd and others.

International collaborations

- IMI RADAR-CNS programme
- TRANSForm
- EU H2020 Developing Human Connectome Project (ERC Synergy Grant)

London School of Hygiene and Tropical Medicine

Top-level insights

- London School of Hygiene and Tropical Medicine (LSHTM) is a world-leading centre for research in public and global health with particular expertise in the optimal application and use of large scale data to answer key questions of public health and clinical importance in non-communicable diseases, infection, drugs, vaccines and health services research.
- They are key partners in the Farr@ London and MRC eMedLab Medical Bioinformatics research investments.
- LSHTM is a centre of excellence in post-graduate training in quantitative methods, with plans to further develop training and education in health informatics.
- It has productive and valuable collaborative links with UK partners that include universities, NHS, industry, data providers and health policy makers and strong; unrivalled international collaborations in population health research.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- LSHTM are key partners in:
 - the **Farr@London** – *Professor Liam Smeeth* is the Deputy Director, with responsibility for leading the biostatistics and training themes
 - **MRC MedBio eMedLab** – *Professor Liam Smeeth* is co-investigator on this award.

Knowledge/Expertise

- LSHTM is a leading innovative centre for biostatistics and applied methodology in the use of large scale health data with a unique concentration of methodologists using state-of-the-art developments in causal inference and disease modelling able to answer key questions of public health and clinical importance.
- It is also a centre of excellence in post-graduate training in quantitative methods, with plans to further develop training and education in health informatics.

Data

- Through their unique collaborations, LSHTM have access to large scale data sources including:
 - Access to UK **National Cancer Registration records** through PHE. For example, they (*Bhaskaran*) carried out the largest ever study on the consequences and obesity for cancer risk using EHR data for 5.24MUK adults.
 - Through the **CONCORD study** (*Coleman*), they have access to individual data for over 25M cancer patients from 279 population-based registries in 67 countries.
 - Access to large-scale data on asthma and eczema through the **Global Asthma Network survey** (*Pearce and Langan*), which involves 2M children from 100 countries.

- Working in partnership with Brazilian researchers, LSHTM (*Smeeth*) is establishing a cohort of over 100M people, built around the **CADASTRO-UNICO** database of EHR from Brazilian social protection programmes applicants. This data will be hosted in a secure datacentre and linked to birth, death, hospital episode, infectious disease notification and other records.
- LHSTM (*Zaba*) have access to the **Longitudinal Population-based HIV data in Africa** network, which links community-based studies in six countries in Africa with high HIV prevalence. Combined data creates an individual-level database with over 6M person-years of exposure time to risk of HIV infection and 140,000 person- years of information about those known to be infected.
- They also have access to data form large international clinical trials data such as the CRASH-1 (over 10,000 patients) and 2 (over 20,000 patients) trials.

Translational and innovation capabilities

Industry links

- Their partnership with AstraZeneca involves collaborations with the Statistical Innovation Group on projects using machine learning to improve risk prediction models and handling missing data in study designs common in EHR.

Policy-making bodies links

- Their methodology research has policy-relevant applications, making LSHTM a world leading centre for shaping vaccine and drug policy. In close collaboration with PHE, they evaluate new vaccines and schedules using novel methods applied to EHR, infectious disease modelling and cost-effectiveness analyses. This work feeds directly into policy-making through membership of and engagement with policy-making bodies, such as the JCVI and Scientific Advisory Group for Emergencies in the UK, and the World Health Organisation (WHO). For example, they use cutting edge methodology to: better understand and predict the spread of infections and inform control policies such as during the Ebola Crisis (*Edmunds*); and predict the impact and cost-effectiveness of changes to the UK national influenza immunisation policy with results adopted by the Joint Committee on Vaccines and Immunisation (JCVI).

NHS links

- LSHTM is part of the UCL Partners (AHSN) and the UCLP AHSC. UCLP Informatics activities are supported by a single informatics strategy across all Partners (*see also section on UCL AHSN*).

Opportunities

- Their extensive network of collaborations, which provides access to a wealth of data sets in combination with their quantitative expertise provide fertile ground for research into innovative methodology development.

Potential gaps

- LSHTM does not seem to have significant research activity in the citizen-driven/digital health area; and public involvement/engagement relevant to health data science.

Other relevant capabilities

Research themes

Discovery science

- Use computational biology and applied genomics to provide insights into pathogens and host interactions, and improve understanding of infectious diseases. For example, work on the characterisation of the genomic variation across thousands of *Mycobacterium tuberculosis* samples to establish a set of SNPs that can be used to barcode sub-lineages (*Clarke*).

Population and public health

- Conducting pioneering work on the use of EHR to conduct clinical trials using EHR data (*Smeeth*).

Disease areas

- A number of areas with particular focus on non-communicable disease (NCD) research and in particular cancer and asthma and allergies; and infectious diseases.

Analytics

- LSHTM is a leader in developing and fitting complex mechanistic mathematical models using cutting-edge Bayesian inferential techniques, used for example to better understand and predict the spread of infections and inform control policies.
- They (*Carpenter*) have expertise in embedding the power of machine learning techniques for handling missing data, with a rich contextual understanding of the unique challenges of routinely collected data, including imputation of missing data in EHR and data from randomised controlled trials; and for handling confounding to estimate treatment effectiveness from large observational datasets.

e-Infrastructure

Compute and Trusted Research Environments

- LSHTM has access to the MRC eMedLab private cloud compute and data storage cluster as a consortium partner. The compute cluster has 6,000 cores and 4.3PB bulk storage that can be configured as bespoke virtual servers to process large bioinformatics datasets.
- For extra capacity to accommodate multi-TB/PB datasets, researchers can request space on the **Storage-on-demand server**.
- LHSTM researchers working with sensitive data register with **LSHTM's secure server**, which uses user authentication and encryption to protect data from unauthorised access. Access to the Secure Server is restricted to LSHTM users only within the campus.

Training

- LSHTM is a national school for public health research and provides outstanding research training opportunities. They have extensive MSc programmes and specialist short courses offer training in advanced quantitative methods; health policy; economic analysis of health interventions; and infectious disease modelling. The MScs in Medical Statistics, Epidemiology and Public Health offer specialised training, producing researchers with the ability to use health informatics data to inform policy.
- LSHTM have had considerable success with fellowship schemes, with 40 fellows currently working in health informatics. This includes 17 MRC Skills Development and Strategic Skills Fellowships on mathematical modelling of infectious diseases, clinical trials methodology and EHR research.

Public Involvement/Engagement

- LSHTM does not seem to have any significant public involvement/engagement activity relating to health data research.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@London (led and based in UCL) – award partners include: LSHTM, QMUL, MRC CTU and the Health Protection Agency
- MRC Medical Bioinformatics eMedLab – partnership between Crick, KCL, LSHTM, QMUL, UCL, LSHTM, EMBL-EBI and Sanger

National data providers

- Public Health England
- CPRD

Industry

- AstraZeneca

International collaborations

- Collaborations with over 100 countries throughout the world, including institutions in Africa, Asia and Latin America.
- World Health Organisation

Queen Mary University of London

Top-level insights

- Queen Mary University of London (QMUL) leads national analytical efforts on stratified medicine using the MRC eMedLab cloud infrastructure which represents a strong collaboration between this MRC Medical Bioinformatics award, the MRC Stratified Medicine consortia and the Farr Institute (mainly UCL and Manchester).
- QMUL has access to large diverse and health disadvantaged populations based in East London served by their partners in: Barts Health NHS Trust (UK's largest hospital seeing 1.4m outpatients/year) and East London Foundation NHS Trust; local primary care and CCGs; and the Tower Hamlets Council. This has been a major resource, which they are using for research and quality improvement programmes.
- This access to large real-time primary and hospital care data is complemented by successfully onward linkage to a wealth of other data including genomics, imaging and geography and further context and expertise for turning the data into meaningful outputs for both local providers and third parties. This has allowed them to develop a model of a local LHS which is now being taken up by others.

Strengths, Capabilities and Opportunities

Strengths

Informatics investment

- QMUL are key partners in:
 - the **Farr@London** – *Caulfield, Petersen, Robson and Timmis* co-investigators on this award.
 - the **MRC eMedLab** – *Caulfield* is co-investigator on this award. QMUL has a significant role in this consortium, providing governance and operational support for the facility: *Caulfield* and *Barnes* are Board Members; *Barnes* is Deputy Chair of the Executive Board and Chair of Technical Governance group; *Christie* leads several aspects of the technical implementation.

Knowledge/Expertise

- QMUL (*Barnes*) provides national informatics leadership on data analysis and integration activities in a number of **MRC Stratified Medicine consortia** (e.g. RA-MAP, MATURA, PSORT); and the establishment of a stratified medicine infrastructure theme with the MRC eMedLab.
- They have strong expertise in **linking across EHR data, geographic data, human genetics, genomics and imaging data** as demonstrated by Genomics England (*Caulfield, Turnbull*), East London Genes and Health (*van Heel*), primary care Clinical Effectiveness Group (*Robson*), Barts Bioresource (*Timmis, Petersen, Caulfield, Barnes*) and the UK Biobank cardiovascular imaging programme activities (*Petersen*).

Data

- They have access to large diverse and health disadvantaged populations of East London through their partners: Barts Health NHS Trust (UK's largest hospital seeing 1.4m outpatients/year); East London Foundation NHS Trust; local primary care and CCGs; and Tower Hamlets Council.

- The **Barts Bioresource** has e-consent for research and recall of cardiovascular patients (11,171 enrolled).
- QMUL also leads in **UK Biobank's cardiac magnetic resonance imaging data** collection and analysis (Petersen) with collaborators at Oxford, Imperial, EMBL-EBI and Circle Cardiovascular Imaging Inc.

Compute

- The current **QMUL HPC facility** is about 5,000 cores available over 250 nodes, in a heterogeneous architecture suitable for a range of applications.
- As mentioned above, they have shared ownership of the MRC **eMedLab** private cloud compute and storage cluster.
- QMUL is one of seven university partners leading, the **Joint Academic Data Science Endeavour**⁷ (JADE, EPSRC £3M award), a national 5 Petaflop GPU facility (one of the largest in the UK) that will support multi-disciplinary science with a focus on machine learning (partly in association with the Alan Turing Institute), image/video/audio analysis, and molecular dynamics. Applications will be in areas including Natural Language Understanding, Autonomous Intelligent Machines, Medical Imaging,

Translational and innovation capabilities

NHS links

- QMUL hosts the **Genomics England's** headquarters with 20 FTE scientists employed by QMUL. They have also received an MRC capital award for the Genomics England safe haven datacentre (£24M).
- The Genomics England Bioinformatics team is forging a global lead in the assembly, analysis and interpretation of genome sequence data, working with a coalition of 2,500 researchers and clinicians assembled for genome/phenotype clinical interpretation (GeCIPs).
- Their **NIHR Cardiovascular BRC** (CVBRC, £6.3M) is a partnership between QMUL and Barts Health NHS Trust, serving 6M people across North East London, exploiting real-world health data and hosting the National Institute for CV Outcomes Research (NICOR) to embed novel trial designs for new devices and interventions.
- QMUL is part of the UCL Partners (AHSN) and the UCLP AHSC. UCLP Informatics activities are supported by a single informatics strategy across all Partners (*see also section on UCL AHSN*).

Opportunities

- QMUL researchers are developing near real-time linkage capabilities of primary (EMIS) and secondary care (Cerner Millennium) NHS EHR data on more than 1.5M people on behalf of local CCGs and NHS Trusts (the **Discovery programme**). This model of LHS with local ownership provides a methodology which could be potentially nationally scalable to other health economies (*see also section on LHS*).
- In the longer term, a major **new Centre for Life Sciences** in Whitechapel, (announced in Jan 2017) will greatly increase partnership opportunities especially with industry, including pharma, digital health and AI.

7. This proposal led by the University of Oxford, with support from the Alan Turing Institute (ATI), Bristol, Edinburgh, KCL, QMUL, Sheffield, Southampton and UCL is for a national GPU system. The architecture is based on "fat" GPU compute nodes, with 8 of NVIDIA's new Pascal GPUs

Potential gaps

- Apart from the ESRC-funded PAMBAYESIAN programme, QMUL does not seem to lead on any significant citizen-driven/digital health research initiatives.

Other relevant capabilities

Research themes

Discovery science

- Human knock-out identification in healthier adults with related parents – they have been actively sequencing cohorts from Born In Bradford, East London Genes & Health and UK Biobank in collaboration with other large international efforts (Broad).
- Genomics (e.g. GeL and East London Genes & Health)

Precision medicine

- QMUL (Barton) co-leads with the University of Manchester (Pitzalis) – the **MRC Stratified Medicine Consortium MATURA** (£5M), which aims to enable early, effective treatments and improve the cost-effectiveness of care for around 500,000 people in the UK who suffer from rheumatoid arthritis.
- In addition, QMUL researchers are **co-investigators** in the following **MRC Stratified Medicine Consortia**: AIM-HY (biological informative markers for stratification of hypertension); and **STOP-HCV** (stratified medicine to optimise treatment for hepatitis C virus infection).

Population and public health

- QMUL's **Wolfson Institute of Preventative Medicine** focuses on translation of epidemiology research into public health strategies and practical implementation (*Cuizic, Sasieni*). Other areas of informatics and analytical excellence include: bridging of interventional and preventative approaches (*Wald*), screening and prevention (*Duffy*), social and community epidemiology (*Priebe*).

Learning Health Systems

- The **Clinical Effectiveness Group** in the QMUL Centre for Primary Care has supported quality improvement and research programmes for over 20 years using EHR data from 1000 GPs who serve the local population (Robson). This Group is also collaborating with two local NHS Trusts, four GP practices, a charity and local authorities on the **Discovery programme**, which has started near real-time linkage of local primary (EMIS) and secondary care (Cerner Millennium) EHR data on over 1.5M people in East London. This will not only enable large real-time linkage but also access to interpretable and accessible outputs for both local providers and third parties providing a model of a LHS with local ownership and a methodology nationally scalable to other health economies.
- Currently, they are demonstrating feasibility exemplar LHS for asthma providing near real-time benchmarking data to GP teams with on-screen decision support to improve process and outcome.

Disease areas

- QMUL's research focuses on a number of areas and in particular on: cardiovascular (key strategic priority), cancer, inherited disorders, rare diseases, RA, asthma, diabetes and obesity.

Analytics

- Using a tranSMART/i2b2 multi-omic data architecture within the MRC eMedLab, QMUL leads the analysis efforts for a number of MRC Stratified Medicine Consortia; and the use of this set up is now extended to other consortia.
- Analysis teams at QMUL have applied an internationally competitive analysis strategy to meet the challenge of the **global open access data policy of UK Biobank**.
- QMUL investigators lead the development of a **UK-Biobank analysis hub environment** on MRC eMedLab to perform analysis on GWAS ((70M variants in 500K subjects)) and CMRI data
- Under our **QMUL Life Sciences Initiative**, interdisciplinary links (e.g. with Geography) are fostered, to encompass analysis of geographic health inequalities and Bayesian disease mapping, including innovative use of a wide variety of data (e.g. ONS census data, PHE Local Indicators and service access indicators).
- Through the **Barts Cancer Institute**, they have access to range of web resources for tissue banking and data mining including: Breast Cancer Now Tissue Bank bioinformatics portal (BCNTBbp); the Pancreatic Cancer Research Fund Tissue Bank (PCRFTB); the Pancreatic Expression Database (PED); and the in vivo breast cancer model database(SEARCHBreast).

e-Infrastructure

Compute and Trusted Research Environments

- The current QMUL high performance computing facility has about 5,000 cores available over 250 nodes, in a heterogeneous architecture suitable for a range of application. The HPC is supported by a 1.1 PB GPFS parallel filesystem, and within the cluster there are 512 GB fat nodes designed specifically for bioinformatics workloads.
- As mentioned previously, QMUL has access to the MRC eMedLab private cloud compute and data storage cluster as a consortium partner. The compute cluster has 6,000 cores and 4.3PB bulk storage that can be configured as bespoke virtual servers to process large bioinformatics datasets.
- They have also built and configured DataCentre, using Aimes secure platform, and data safe haven for the purposes of the Discovery programme.
- The Barts Cancer Institute has also an identified clinical data safe haven.

Training

- QMUL has an integrated cross-faculty suite of training and capacity building programs around health and biomedical informatics, including MSc programmes in Bioinformatics, and Medical Informatics (to start in 2018).
- Our LSI Centre for Computational Biology delivers world-leading training in Computational Biology, Bioinformatics and Medical Informatics to undergraduates, Masters and PhD students, and through CPD courses.
- They also co-lead the Farr London Academy, with PhD programme (QMUL hosts 2 PhDs), MSc programmes and short courses.

Public Involvement/Engagement

- QMUL have not mentioned any specific PPI/E activities in their Sol.

Partnerships

- QMUL has close partnerships with a large number informatics investments and stakeholders which are mentioned in the sections above.

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@London (led and based in UCL) – award partners include: LSHTM, QMUL, MRC CTU and the Health Protection Agency
- MRC eMedLab – partnership between Crick, KCL, LSHTM, QMUL, UCL, LSHTM, EMBL-EBI and Sanger
- MRC Stratified Medicine Consortia
- Genomics partnerships with Oxford (Wellcome Trust Centre for Human Genetics) and the Wellcome Trust Sanger Institute
- Cardiovascular imaging and UK Biobank partnerships with Imperial

DH and NHS

- NIHR Barts Biomedical Research Centre
- Barts Health NHS Trust and Homerton University Hospital NHS Trust
- North Thames CLARHC – support quality improvements projects
- UCLP AHSN –support the East London model of utilisation and quality improvement
- CCGs

Data providers

- Public Health England
- ONS
- EMIS

Industry

- Genomics England (QMUL) partnerships with 12 companies
- QMUL and Quintiles IMS transnational partnership.
- Technology companies: CIRCLE Inc. and Clinithink's

University College London

Top-level insights

- The University College London (UCL) strengths in health and biomedical informatics build on the Farr@London and MRC eMedLab activities, which together put in place significant critical mass of excellent health informatics and data science expertise (954 academics and research staff including 22 senior faculty fellows based at Farr@London); and resources (£1,000M of existing external research grants in data science and informatics with access to 24k cores and 8 PB storage).
- UCL has a strong interdisciplinary environment, including engineering, social sciences, health psychology (including digital behaviour change), medical humanities and law.
- In addition, there is a strong translational pipeline, including informatics research, and links to NHS: three BRCs, an NIHR HIC (critical care) site, an AHSC, and an AHSN, including 23 NHS Trusts covering a diverse population of 6M.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- UCL are leading the activities of the Farr Institute in London (Professor Hemingway is the Director of the Farr@London) (see also section 2.1.1).
- In addition, UCL (*Luscombe*) have been at the forefront of the MRC eMedLab activities (see also section 2.1.2).
- The UCL NIHR BRC is one of the five BRCs involved in the NIHR **HIC programme** (see section 2.2.2), leading on critical care research theme.
- UCL has a top-ranked **UK Computer Science department** and the associated world-leading **Centre for Medical Image Computing (CMIC)**.

Knowledge/Expertise

- The collective informatics research talent possessed at UCL ranges from computational science to handling and interrogation of large cohort data sets, bioinformatics and decision support tools.
- The **UCL Institute of Digital Health, ESRC Administrative Data Research Centre (ADRC)** – England, and the **UCL Institute of Healthcare Engineering** add a **strong interdisciplinary environment**, including engineering, social sciences, health psychology (including digital behaviour change), medical humanities and law.
- Furthermore, they are hosts to an array of leading institutes such as the **UCL Institute of Clinical Trials and Methodology** (the largest collection of trialists in Europe ~400 staff, incorporating the MRC Clinical Trials Unit, covering phase 1-4 trials); imaging including **CMIC**, the **EPSRC/Wellcome Trust Imaging Centre**, and **EPSRC iSense**.

Data

- UCL hosts numerous clinical cohorts and has access to data with great scale and phenotypic and genotypic depth that include:
 - the 10M population in **CALIBER**, with large scale linked EHR (with phenotyping and other tools of scalable, replicable research);
 - 500K people (UK Biobank, with automated pipelines for GWAS and PheWAS);
 - UKCTOCS – 180K women (with serial serum measures);
 - 30K international cases in GWAS collections (e.g. HERMES for heart failure);
 - 14 population-based UK cohorts including those focused on early life influences (the 4 UK national birth cohorts~58k total), multiple repeat measures in mid and older age (Whitehall II, British Regional Heart Study, British Women’s Heart Study), ethnic minorities (SABRE), and older people (ELSA).
- Through the **NIHR HIC**, UCL hosts critical care data. This dataset contains identifiable data supplied by all HIC participating Trusts, and stored in the UCL Trusted Research Environment (TRE). It is secondary care, deep episodic patient-level data including: demographic, physiological, laboratory and drug data for all patients admitted to Critical Care and Intensive Care Units from January 2014 onwards (c. 8,000 - 10,000 patients).

Translational and innovation capabilities

Industry links

- Industry partnerships in data science include collaborations with **pharmaceutical** (GSK, AstraZeneca and Novartis) and **technology companies** (Intel, Cerner, Google DeepMind Health).
- UCL has also pipelines with SMEs. An early success is the HELICON Health, a digital health SME spun out of UCL, which is one of the 32 companies that make up the 2016-17 cohort of the **Digital.HealthLondon Accelerator** programme.
- In addition, access to £40M in **venture capital** is available to support to UCL’s translational technology.

NHS links

NIHR Biomedical Research Centre

- The **UCL Hospitals BRC** (£111.5M; *Director: Professor Bryan Williams*) is a partnership between UCL and the University College London Hospitals NHS Foundation Trust. The BRC has a cross-cutting theme on Healthcare Informatics, Genomics/’omics, Data Science (*Hemingway*).

It is the only BRC in England with the Farr Institute of Health Informatics Research, the Crick Institute, the Alan Turing Institute (computer science), and one of the world’s largest biomedical cloud infrastructures (MRC eMedLab). Its activities are aligned with these of Farr@London.

There is also a **UCL BioResource** – a panel of over 7,000 volunteers who have agreed to be approached to participate in future medical research. Both volunteers and patients are represented and the size of the panel is always growing, with an aim to reach 10,000 by the end of March 2017.

- In the **NIHR HIC** this BRC leads the critical care research theme (*Singer*).

Academic Health Sciences Centre

- UCL is part of the **UCL Partners Academic Health Science Centre (AHSC)** (*Academic Director: Professor Sir John Tooke*), which comprise three HEI:

1. London School of Hygiene and Tropical Medicine
2. Queen Mary University of London
3. UCL

and five NHS trusts:

1. Barts Health NHS Foundation Trust
2. Great Ormond Street Hospital for Children NHS Foundation Trust
3. Moorfields Eye Hospital NHS Foundation Trust
4. Royal Free London NHS Foundation Trust
5. University College London Hospitals NHS Foundation Trust

It is the largest AHSC in Europe support a population of 6M people and over 40 health, social care and academic organisations.

The collective informatics research talent possessed by the partnership ranging from computational science, computational modelling, handling and interrogation of large cohort data sets, bioinformatics, decision support tools, m- and e-health, preventative, diagnostic and therapeutic initiatives has a physical focus in the Farr Institute @ London.

Academic Health Sciences Network

- UCL is also part of the **UCL Partners Academic Health Science Network**.

Informatics activities are supported by a single **UCLP informatics strategy** (2014-2017) across all Partners; and a single Informatics Board with an Independent Chair (*Professor Andrew Morris*) and a small Informatics Programme Executive Team, who will be predominantly embedded in local health economies and will act in an enabling capacity to support delivery of the Strategy.

- The UCLP AHSC, with **Med City** and the other AHSCs and AHSNs in London are also part of the **DigitalHealth.London** initiative.

Collaboration for Leadership in Applied Health Research and Care

- Through partnership with the **NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames**, they are involved in the evaluation of innovations that include novel digital tech and determinants of their implementation.

NHS 100,00 genomes project | Genomics England Ltd

- UCL contributes to one of the largest Genomics England Genomics Medicine Centres and lead on the rare diseases theme. Led by GOSH, six NHS trusts in north London have formed the **North Thames Genomics Medicine Centre**.

Opportunities

- With the NHS and other key delivery partners UCL *plans* to establish a **trusted research environment**, based on UCL Partners population as part of the UK federated data ecosystem. This will deliver data integration across primary, secondary and social care (building on existing London examples in mental health) and across data types including health records, imaging, genomics, and mobile and wider societal data. This safe and secure environment will allow investigation with: identifiable data of data driven solutions, including real time decision support, to immediate NHS problems (e.g. cause specific readmission); and with pseudonymised data used by researchers to deliver theme research priorities and by innovation partners to grow the economy.
- Plans to create a new 25 FTE **Clinical Research Informatics Department** with UCLH as part of the NIHR BRC Health Informatics theme (£7.5M) building on their HIC project.
- UCLH has just bought Epic and NHS is offering a £20M incentive to collaborate with University of Cambridge that was the first Trust to use it and develop a research gateway for this resource.

Potential gaps

- UCL is not leading on any of the MRC Stratified Medicine Consortia and does not appear to be involved in any other stratified medicine initiatives.

Other relevant capabilities

Research themes

Discovery science

- UCL has strengths in using data science approaches to redefine diseases in mechanistic, treatment-relevant taxonomies. Examples include:
 - the **ICONIC project** (funded by DH and the Wellcome Trust and UCL-led) has developed sustainable next generation technology for the viral full-length genome sequencing using residual diagnostic samples within the diagnostic value chain. ICONIC processes and assembles de novo the virus genomes using high performance computation available in ‘cloud resources’ bypassing the need for embedding complex computer technology in the NHS and specialist staff.

Precision medicine

- UCL's and eMedLab have contributed integrating infrastructure to some of the MRC Stratified Medicine Consortia
- They have established an interdisciplinary **Precision Medicine Accelerator** for nurturing start-ups.

Population and Public health

- One of UCL's major focus is on trials embedded in clinical practice and in particular, the use of data science approaches for the development of novel adaptive designs and identification of novel patient groups
- Using the CALIBER platform, UCL researchers have linked primary and secondary care data on approximately 2M people aged more than 30 year, and free from diagnosed cardiovascular disease, to understand cardiovascular outcomes and presentations. A similar approach has been taken for other diseases, including diabetes and cancer.
- Farr@London investigators (*Hemingway, Denaxas, Dobson*) are leading the **EU IMI BigData@Heart** consortium (€20M). This research programme will integrate healthcare data, wearables, genetic profiles, information about patients' lifestyles and health and their own reporting of symptoms, to better understand causes of different CVD conditions and improve patient care.

Citizen-driven/digital health

- Through the **NIHR TB Reach programme**, UCL researchers are leading the world's first randomised controlled trial of Video Observed Therapy (VOT) to support patients to complete tuberculosis treatment in the UK using a smartphone app developed by the University of San Diego.

Learning Health Systems

- The **ICONIC** project (also mentioned above) links virus genetics and diagnostic information to hospital based clinical information and generates the appropriate clinical reporting tools. This method allows researchers to identify links between cases that would otherwise have been missed, target infection control processes to specific wards and determine whether sudden surges in case numbers reflect several smaller clusters of infection or a single uncontrolled outbreak. Hospital teams across the UK are piloting these methods and evaluating how hospital flu can be better managed to improve patient outcomes and save costs.

Disease areas

- UCL has research programmes in a number of disease areas, and in particular: cancer, cardiovascular, neuro-degeneration and mother and child health.

Analytics

- **CALIBER** has become a leading research platform for the understanding of health progression, from onset of early stages to the development of significant functional impairment. The platform links electronic health records from primary care, coded hospital records on diagnoses and operations, social deprivation information and cause-specific mortality data in England for roughly 10M adults. An online Data Portal contains disease and risk factor phenotyping algorithms for the wider scientific community to use.
- In collaboration with the **ADRC–England**, Farr@London investigators have developed methods for mother-child linkage of 10M deliveries in England and are working with Public Health England and the Ministry of Justice to apply new linkage methods to government-held data.
- Farr@London investigators are developing **decision support tools** for a variety of settings and clinical conditions, including tools to: investigate variations in tonsillectomy rates and recurrent sore throat infection for patients undergoing the procedure; and to predict risk of multi-drug resistant tuberculosis based on epidemiological and clinical data linked to notification and lab data; identify tuberculosis from serial chest radiographs.
- **Google DeepMind Health** has partnered with the UCL (Rees) and the Royal Free NHS Foundation Trust on their Streams app to support clinical decisions for prevention of acute kidney injury.
- They are also working with Google and **Microsoft Research** to investigate the use of search engine logs and twitter to monitor influenza occurrence and secondary household transmission rates validating methodology against FluWatch community cohort.

e-Infrastructure

Compute

- Access to 8 PB storage and 24,000 cores.
- There is also access to up to 2 PB high performance tiered storage with a 10GBps connection to Legion compute cluster (Farr-funded total cluster size 10,000 cores).
- UCL has access to the MRC eMedLab private cloud compute and data storage cluster as a consortium partner. The compute cluster has 6,000 cores and 4.3PB bulk storage that can be configured as bespoke virtual servers to process large bioinformatics datasets.

Trusted Research Environment

- UCL has established an ISO27001 Data Safe Haven and associated IG framework in 2012 which was further enhanced through the Farr Capital funding. It currently supports a total of 322 projects with 765 users. There is an associate 70 TB tiered storage.

Training

- UCL has developed a significant programme of training and capacity building to create future health care data science leaders, including an innovative **mini-MD clinical immersion course** that delivers intensive exposure of non-clinicians to all elements of a patient pathway used successfully in medical imaging and cancer.
- **Ten MSc programmes** provide core disciplinary expertise totalling 366 students in 2016/17.
- UCL Computer Science Partners with Google DeepMind to deliver an **industrial partnership PhD programme in Machine Learning**.
- Currently developing an intercalated BSc Big Data for Health for medical students, an MSc in Data Analytics and NIHR Clinician Coders (for NIHR and NHS trainees).

Public Involvement/Engagement

- Through Farr@London, UCL has been hosting a number public involvement/engagement activities related to health data science. For example, workshops for researchers and lay members of the public to deepen understanding of the reasons for, and impact of, missing information in EHR from patient and researcher perspectives. Also, it has been leading on scientific engagement with schools, promoting the idea that students benefit from performing real science using real data in the classroom. Findings from Decipher mydata!Flu! based on data from 27 schools show that school sickness absence data are correlated with influenza surveillance data and may provide a good warning system for spotting flu outbreaks.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@London (led and based in UCL) – award partners include: LSHTM, QMUL, MRC CTU and the Health Protection Agency. Through this investment, strong links to the other Farr Centres: Farr@Scotland, Farr@HeRC (Manchester) and Farr@CIPHER(Swansea)
- MRC eMedLab – partnership between Crick, LSHTM, KCL, QMUL, UCL, EMBL-EBI and Sanger
- MRC Stratified Medicine Consortia
- UK Biobank – collaboration on the adjudication of outcomes through EHR phenotyping
- ESCRC ADRC – England and the other ADRCs and ADRN
- Imperial College London
- Francis Crick Institute
- Alan Turing Institute – founders are the Universities of: Cambridge, Edinburgh, Oxford, UCL, Warwick
- National Institute of Cardiovascular Outcomes Research
- UK Dementia Research Institute

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:

DH and NHS

- NIHR BRCs – UCLH, GOSH and Moorfields
- NIHR HIC
- UCLP AHSC and AHSN including 23 NHS Trusts – and also, strong links with the other two London AHSC and AHSN
- Genomics England – hosts a GMC
- North Thames CLARHC – evaluation of innovative digital health tech

Data providers

- Public Health England
- ONS
- CPRD

Industry

- Pharma companies: GSK, AstraZeneca, Novartis, Roche
- Technology companies: Intel, Cerner, Google DeepMind Health, Microsoft Research
- Biotech: Illumina, Somalogic

International

- IMI: Big Data@Heart; EHR4CR; and EMIF
- P-medicine
- ContraCancrum
- Virtual Physiological Human Network
- IT Future of Medicine initiative
- Karolinska Institute – collaborative projects
- Ontario Institute for Clinical Evaluative Sciences – collaborative projects
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR

GEOGRAPHIC REGION: ENGLAND – NORTH

- Statements of Interest (Sols) to become a substantive HDR UK site were received from the following five ROs based in the North of England:
 1. **University of Leeds**
 2. **University of Liverpool**
 3. **University of Manchester**
 4. **University of Newcastle**
 5. **University of Sheffield**
- Due to time/resource constraints, this report contains information on ROs which submitted Sols for HDR UK Substantive Sites only. The University of Lancaster and the University of York, which are based in this geographic region, had also submitted Sols to become a Partner HDR UK but this report doesn't contain any information on their informatics/health data research investments.

Cross-region Strategic Activities

- It is noteworthy that all of the ROs already mentioned in this geographic cluster, as well as the University of Durham, constitute what is known as the **N8 Research Partnership**. This is a collaboration of the eight most research intensive Universities in the North of England that supports the establishment of multi-partner bio-economy and preclinical research programmes.

Region-wide health and biomedical informatics activities

- The **N8 HPC** is an EPSRC-funded project to provide Tier2 HPC facility to the N8 Research Partnership (with a total of 332 compute nodes (5,312 cores); and 174 TBytes Lustre v2 parallel file system for storage.
- The ESRC-funded **Centre for Doctoral Training (CDT) in Data Analytics and Society** (£6M) is led by the University of Leeds in collaboration with the Universities of Liverpool, Manchester and Sheffield. This investment is intended to pump prime the development of training in new, and emerging areas, and will focus on: real-time data analysis; data linkage; and interoperability of data from different sources. The substantive focus is on '**new forms of data**', in a broad sense, and on the opportunities provided by these data to study social processes as they unfold. Funding is for three cohorts of at least 10 students per year. The CDT will start taking students in October 2017.

Other region-wide healthcare research related activities

- The N8 Universities are also members of the **North Health Science Alliance** (NHTSA) that supports the establishment of multi-centre clinical research and development projects. Since its formation in 2011, the NHTSA has worked to be a voice for the North's health science community joining together, at-scale and at-pace, capabilities and resources across eight University Medical Schools, eight research intensive NHS Trusts and four Academic Health Science Networks across North England (15M population).

Its aim is to enable companies to grow and create jobs by helping the NHS and industry engage more productively, focused on proven innovations that meet defined local needs.

So far, NHSA has: brought in £60M worth of contracts to the North; acted as a collaborative voice for the North's £10.8bn health science economy; worked with international commercial companies; and led successful collaborations to attract health science investment.

- One such investment is the world-leading **Health North: Connected Health Cities** programme funded by the Department of Health and Social Care (£20M) in 2015. Health North is operating across North England to generate innovations that can deliver more effective and efficient health and social care. Its foundations are to:
 - establish a social contract with citizens that gives license to use health data for public good;
 - produce timely and actionable information from patient and population data;
 - use new intelligence sources to understand pathways of care across different provider organisations and to target resources to needs in much more agile and specific ways than at present; and
 - accelerate business growth in the digital health sector in North England.

In its first phase, Connected Health Cities initiative will be piloted in four city regions (1. Greater Manchester; 2. Yorkshire; 3. North West Coast; and 4. North East & North Cumbria), with the common principles of:

- engaging and involving the public to build public trust and civic partnerships;
- working with data custodians and existing infrastructure at local, regional and national level to create linked-data critical masses for deeper understanding of health in defined populations;
- developing at the heart of each city region an Ark – a secure, combinatorial data analytics facility, with state-of-the-art data management and analysis tools, underpinned by research, education and training;
- bringing people from academia, NHS and industry into deep collaboration in the Ark, working in partnership to create new knowledge that will inform decision-making at levels.

The aim is to pilot data-intensive health service optimisation methodology and understand how it can work efficiently in city regions, then to network those regions so they can 'borrow strength'.

- The Newcastle University-hosted **GCHQ Academic Centre of Excellence in Cyber Security Research** is the founding member of **CyberNorth**, an initiative to improve cyber security practices in the North of England and to create entrepreneurial and other activities in cyber security that benefits the economy of the region.

University of Leeds

Top-level insights

- The University of Leeds strengths in health and biomedical informatics build on the MRC Medical Bioinformatics award, which supported the establishment Leeds Medical Bioinformatics Centre. This award has been pivotal in the founding of the Leeds Institute of Data Analytics (LIDA), which also works in combination with the ESRC Consumer Data Research Centre (CDRC) to provide a network for data analytics in the University.
- LIDA itself provides a unique collaborative working environment currently hosting colleagues from several private and public sector organisations, including Leeds City Council, the three Clinical Commissioning Groups within Leeds, Call Credit and Quintiles IMS. This model of collaboration through co-location has proven to be an effective means of directing research capability towards external need, thereby driving socio-economic impact.
- Furthermore, strong local partnerships between the University of Leeds and health and social care providers (including the Leeds Teaching Hospitals NHS Trust, NHS Digital, TTP and EMIS) has enabled the development of the Leeds Care Record, an EPR system that spans primary, secondary and social care.
- Health and biomedical data science training across the career pathway from undergraduate to early career researchers and clinical academics. They are also using their interdisciplinary structure and approach to improve cross-departmental transition of quantitative researchers through mechanisms like the Data Science Interns and Associate Fellows programmes.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The **Leeds Medical Bioinformatics Centre** (*Director: Professor Sir Alex Markham*) was established through an MRC Medical Bioinformatics award in 2013 (£7M; see also section 2.1.2). The Centre combines methodological strengths in bioinformatics, statistical epidemiology and health data analytics with a multi-disciplinary approach to applied health informatics, consumer data and social data analytics.
- The **Leeds Institute of Data Analytics (LIDA)** (*Director: Professor Mark Birkin*) was established through £40M matched MRC and ESRC funding, with an additional £8M of University spending, to provide a flagship environment for health and biomedical informatics research. It has the capability to link 'omics, phenotypic, imaging and clinical data in a fully accredited and secure environment with the ambition to develop as a centre for personalised medicine and health. Subsequent alignment with associated projects from 34 Schools in the UoLeeds provides leveraged funding of £45M in LIDA.
- The **National Consumer Data Research Centre (CDRC)** (*Co-Director: Professor Mark Birkin*) (ESRC-funded investment) – CDRC holds extensive data sets on diet, health and mobile monitoring and active travel
- The EPSRC-funded **QuantiCode** project aims to deliver an infrastructure that provides far more powerful analytical tools than those available today for public and private sector organisations to transform their abilities to analyse quantitative and coded longitudinal data. It will do this through: understanding data workflows and addressing the barriers of knowledge extraction; leadership in data governance; efficient heterogeneous data fusion; robust and scalable data mining/machine learning tools; data visualization/mining of abstraction models.

Knowledge/Expertise

- University of Leeds has a long track record in clinical trials and applied health research with expertise in developing new methodologies for data linkage, patient selection and consent. This is in combination with data science expertise in mathematics and computing plus diverse inputs from social and environmental science, humanities and business.
- Enhancing interdisciplinary research is a key aim of the University of Leeds with an £1M in salaries and infrastructure invested per annum to support grand challenge-led inter-disciplinary research themes in Health, Water, Food, Energy, Culture, Cities and High Value Engineering.

Data

- University of Leeds has access to large complex datasets from patients and populations that includes rich clinical phenotypic data from EPR. Data from Leeds Teaching Hospitals NHS Trust is now available for research purposes across all of the Trust's systems. For example, data from over 70 clinical and administrative systems, including the patient administration system, the oncology EPR, theatre management, prescribing, A&E, pathology and radiology, have been pulled together and linked in a data warehouse.
- Data is also available from disease-specific national datasets and clinical trial datasets (including tens of thousands of patients and multiple conditions).
- Also, access to a unique combination of datasets relating to: Consumers (ESRC Consumer Data Research Centre) generated by retailers and other service organisations as part of their business process, and licenced for academic use through the CDRC; Environment (NERC National Centre for Atmospheric Science); and Security (N8 Policing Research Partnership).
- This is in addition to imaging, structural biology, NGS, and digital pathology data generated from the University of Leeds's high throughput research facilities.

Compute

- In addition to access to the EPSRC N8 HPC, University of Leeds has supercomputer infrastructure dedicated to LIDA-based researchers (MARC1) and additional resources for the rest of the researchers at the University (ARC1/2).
- The University of Leeds **Integrated Research Campus** is a secure, IG-compliant data platform that provides access to isolated virtual research environments, which enable data analysis through remote access, rather than data being released.

Translational and innovation capabilities

Links with industry

- University of Leeds has collaborations with a number of external partners including industry. For example, LIDA's Intern Programme for ten postgraduate data scientists annually has specifically been open to external partners such as Rolls Royce, Leeds City Council and Morrisons.
- Commercial collaboration has been supported through a variety of co-funding models including a number of KTP programmes funded by Innovate UK.
- Co-location with the ESRC Consumer Data Research Centre has led to new collaborations, sharing data with Sainsbury's, AQL and Consumer Data.

Links with NHS

- Local partnerships reflect University of Leeds's close relationship with health and social care providers in the Leeds area. Interactions between University, NHS Trust, CCGs, Community Health and Social Care are moderated through a coordinated **Leeds Academic Health Partnership**.
- Furthermore, geographic co-location facilitates strong partnerships with the two major suppliers of **primary care IT systems** (TPP and EMIS); and **NHS Digital**.
- These collaborations have supported the development of electronic patient records (e.g.PPM+) in the Leeds Teaching Hospitals Trust; and the **Leeds Care Record**, which is a very large (2.9M patients) and detailed integrated clinical record combining data from **all** primary care, secondary/tertiary hospital care, mental health, social and community care across Leeds.
- As a partnership between Leeds Teaching Hospitals NHS Trust and the University of Leeds, the **NIHR Leeds BRC** is a renowned centre for translational research into individually targeted, patient focussed therapies across musculoskeletal diseases.

Opportunities

- The University of Leeds's exceptional relationships with health data providers (e.g. NHS Digital, TPP and EMIS), manufacturers, retail organisations, government, policy- makers and patient groups provide an important collaborative infrastructure which spans the full range of interest groups in health and biomedical bioinformatics.

Potential gaps

- The University of Leeds doesn't seem to be leading on any significant research programmes in the areas of citizen-driven/digital health and learning health systems analytical methodology.

Other relevant capabilities

Research themes

Discovery Science

- **LeedsOmics** – is a virtual institute that aims to engage and unify the critical mass of 'omics (genomics, transcriptomics, translomics, proteomics, metabolomics, epigenomics) researchers at the UoLeeds. This effort encompasses investments that include:
 - The **MRC Single Cell Functional Genomics Centre** (£1.1M) provides capability to understand intra-tumour heterogeneity in malignant brain cancer among other diseases.
 - **MIMOmics** (Methods for Integrated analysis of Multiple Omics datasets) is a FP7-HEALTH-2012 Innovation project funded by the European Commission (2012 -2017), coordinated by the Leiden University Medical Center with UoL one of the consortium members.
- The new **National Centre for Hyperpolarised MRI** (£6.8M; MRC) based across the Universities of Leeds and York. The aim of the centre is the clinical translation of a hyperpolarised MRI method known as Signal Amplification by Reversible Exchange, or SABRE, which has been developed by scientists at the University of York. In Leeds, the centre is located within Leeds Teaching Hospitals NHS Trust in order to facilitate clinical studies.
- The **Astbury Centre for Structural Molecular Biology** brings together researchers from across the University of Leeds – largely from physics, the biological sciences and chemistry – to allow interdisciplinary approaches to be harnessed to understand the molecular basis of life.

Precision medicine

- The University of Leeds has a **Precision Medicine Catapult Centre of Excellence**, which is designated as a lead “informatics” centre.

Population and Public health

- The **Leeds Institute of Clinical Trials Research** is a leading UK academic clinical trials unit. Its research spans a broad range of diseases and treatments and specialises in the design, conduct and analysis of both early and late phase academic clinical trials. This includes: design and implementation of trials, including registry-based trials; identification of eligibility; characterisation of ‘treatment as usual’; optimising content and method to measure outcomes; access to follow-up data, health economic modelling of cost and cost effectiveness; and statistical modelling of complex healthcare pathways linked to patient outcomes to identify the likely implications of service changes, including disinvestment decisions.
- The **Yorkshire Lung Study** is UK’s largest lung cancer screening trial will take place in Leeds following a £5.2M investment by regional charity Yorkshire Cancer Research. The trial, carried out in partnership with Leeds Teaching Hospitals NHS Trust, will test screening in mobile vans, which will travel to communities in the Leeds South & East and Leeds West CCGs.
- The **Wolfson Centre for Applied Health Research** (£3M) is a new centre to improve the health and well-being of children and the elderly – and the safety of patients in hospitals and clinics. It brings together researchers from the Universities of Leeds and Bradford with clinicians from Bradford Teaching Hospitals NHS Foundation Trust to address key health priorities for the county and host the Born in Bradford and Born in Bradford’s Better Start cohorts.
- The **Leeds Cancer Epidemiology** group provides support to the national colorectal cancer intelligence hub, which is based in LIDA.

Citizen-driven/Digital health

- No details were provided about activities in this theme in the Sol – or could be identified through desk research.

Learning health systems

- No details were provided about activities in this theme in the Sol – or could be identified through desk research.

Disease areas

- The University of Leeds’s main areas of clinical excellence are: cancer, CVD and musculoskeletal.

Analytics

- University of Leeds researchers work with academic and commercial partners to develop new techniques to capture and extract structured data from plain text, digital imaging and digital pathology.

e-Infrastructure

Compute

- HPC resources are provided by the Advanced Research Computing (ARC) Team. A dedicated HPC, **MARC1**, is prioritised for researchers associated with LIDA. This cluster has 57 standard nodes (256GB) and 2 “large memory” nodes (3TB), with 0.5PB of parallel Lustre storage. The ARC team runs two other clusters, **ARC1** and **ARC2**, for use by any Leeds researcher, the Polaris cluster for researchers at any N8 university, and the shared memory SGI UV2 machine owned by the Farr@HeRC (Manchester).

Trusted Research Environment

- The UoLeeds **Integrated Research Campus** (IRC) is a secure data platform that processes a large volume and variety of data for research purposes. Accreditation by ISO27001 Level 2 is expected in March 2017. IG Toolkit Level 3 is available within Leeds Institute for Clinical Trials Research for projects requiring greater security. It now supports 128 projects of which 2/3 are related to biomedical and health informatics. Data is hosted from Leeds Teaching Hospitals NHS Trust Data Warehouse and commercial and government partners (e.g. Sainsbury, Asda, Zoopla, and TPP). The IRC provides isolated virtual research environments (VREs) which enable data analysis through remote access into the VRE, rather than data being released. Researchers must complete IG training and sign an IRC User Agreement to be given access to the VRE. Data cannot enter or leave the environment without approval. The physical architecture of the IRC is based on two VM clusters split across two data centres on our campus and hosted on a total of 24-node clusters with access to a total of 2PB of storage.

Training and Capacity

- University of Leeds has appointed two new professors in Data Analytics and Biostatistics and Spatial Analysis.
- Furthermore, investment in the next generation of research leaders has been made through the University's **Academic Fellowships**, providing new tenure track appointments for 36 positions linked to LIDA (14 of these have research interests directly aligned to the HDR UK). A further cohort of Fellows will be recruited to a five-year research programme with the **Alan Turing Institute**.
- **MSc programmes** in: Statistical Epidemiology; Data Science and Analytics; Consumer Data Analytics; Health Informatics, Data Analytics; Biomedical and Health Care Ethics.
- **Data Scientist Intern** is a unique programme developed by Leeds with ten postgraduates, mentored and hosted in LIDA, multi-disciplinary project mix and personal development portfolio.
- An integrated **PhD in Data Analytics** is delivered via the highly multi-disciplinary **ESRC CDT in Data Analytics and Society Centre** directed by Leeds with the Universities of Liverpool, Manchester and Sheffield.
- The **Clinical Academic Training** is an NIHR programme to train clinicians to understand data science approaches and learn the basic mathematical and epidemiological skills needed to tackle clinical research challenges.
- Embedding the MRC notion of the “research technology specialist”, University of Leeds has established a **Data Services Team** of specialists with PhD, postdoctoral, commercial and public sector experience to provide underpinning technical expertise to projects and programmes. A dual reporting structure through IT Services and LIDA Senior Management provides opportunities for more flexible career development across different pathways.

Public Involvement/Engagement

- Active engagement of patients and the public, tackling ethical issues and data sharing discussions locally and nationally through collaborations with NHS Digital, and the UseMyData project.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- N8 Research Partnership (Universities of Durham, Lancaster, Leeds, Liverpool, Manchester, Newcastle, Sheffield, and York)
- Northern Health Science Alliance
- Edinburgh, Oxford, UCL
- UK Biobank
- Alan Turing Institute

NHS and other government organisations

- NIHR Leeds BRC
- NHS England
- Department of Health and Social Care
- Environment Agency (UK)
- local CCGs
- Leeds City Council

Data providers

- CPRD
- NHS Digital
- PHE
- ResearchOne

Industry

- Pharma: Roche, Caris, QuintilesIMS
- Tech: AQL, EMIS, IBM, TTP

International

- EU MIMOmics Consortium
- Africa (CDC, Rwanda)
- South America

University of Liverpool

Top-level insights

- The University of Liverpool's strengths in health and biomedical informatics focus on the areas of precision medicine, infectious diseases and clinical trials methodology. Through its collaborations with the Farr@HeRC (Manchester) and the North Health Science Alliance (NHSA), it aspires to deliver a learning health system that will improve patient care and public health.
- There are strong partnerships with local NHS Trusts, four of which have been awarded NHS England Global Digital Exemplar status, will provide opportunities to align its joint applied research programmes and answer important clinical and public health questions using new informatics and technological methods.
- The Liverpool City Region has an array of nationally and internationally significant digital and creative assets. Foremost among these is the STFC Hartree Centre at Daresbury, which houses the most powerful supercomputer in the UK dedicated to industrial R&D. The University has a strategic partnership with the Centre, enabling staff to access its computing facilities. In 2015, a new £113M UK government investment in the Hartree Centre was announced, plus £200M from IBM, involving the deployment of its Watson cognitive computing platform.
- The University of Liverpool is located within the Knowledge Quarter, where £1B of investment is creating one of the largest health campuses in Europe. Furthermore, the digital space in Liverpool currently employs 13,600 people, valued at £985M GVA 2015, and is the second fastest growing technology cluster in the UK (TechNation 2015). Liverpool is one of only three Internet of Things (IoT) technology clusters in the UK (Sir Mark Walport's 2014 review).

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The University of **Liverpool Healthcare Data Laboratory**, established as part of the Connected Health Cities programme, links the University's critical mass of analytical expertise with a programme of engagement of front-line clinicians across the North-West Coast region, supporting iterative co-production of clinically-validated analytics and data visualizations for defined patient pathways with NHS staff and stakeholders.
- University of Liverpool is also a partner in the ESRC-funded **Consumer Data Research Centre** (*Deputy Director: Singleton*), which is based in Leeds.
- University of Liverpool hosts the **Liverpool Centre for Mathematics in Healthcare** (£2.5M; EPSRC), which is one of only five such Centres nationwide. Researchers in the Centre (*Chen*) are developing novel mathematical tools that address serious healthcare challenges, carrying out multi-disciplinary research to explore how mathematics and statistics can deliver more refined and accurate predictive models, and tools, for personalised healthcare delivery. For example, they are developing novel 3D variational segmentation models for medical imaging, incorporating incomplete and partially segmented information, in collaboration with an industrial partner.
- The University of Liverpool investigators (*Williamson, Capewell*) have been collaborators on the **Farr@HeRC** in Manchester award from its inception.

Knowledge/Expertise

- University of Liverpool researchers have experience in extracting, managing and linking data from diverse datasets for policy-relevant analytics, including consumer, finance and welfare datasets not routinely used in applied public health research; as well as stratified medicine, including ‘omics and clinical data.
- One of their key strengths is expertise in clinical trials methodology including the development of analytics, methods and standards. For example, they lead the **COMET Initiative** (Core Outcome Measures in Effectiveness Trials) (*Williamson*), which brings together people interested in the development and application of agreed standardised sets of outcomes, known as ‘core outcome sets’.

Data

- With Liverpool CCG, Liverpool City Council and Alder Hey Children's Hospital University of Liverpool (*Taylor-Robinson*) is establishing an **anonymised population-level child health and development dataset** for all children in Liverpool, nested around the universal health check data, linked to birth records, data on use of local health and social care services, educational outcomes, and National Child Monitoring Programme data. This will form the basis of an electronic population cohort to allow epidemiological assessment of the early origins of health inequalities in Liverpool.
- Through a collaboration between the **NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North West Coast**, the ESRC-funded **Consumer Research Data Centre** and 10 local authorities across the North West, University of Liverpool researchers (*Barr*) have established an Integrated Longitudinal Research Resource of linked neighbourhood datasets, enabling the tracking of the determinants of health and health outcomes within neighbourhoods along with novel neighbourhood-level contextual indicators of networked distances to health care facilities, health-related commercial outlets (fast food, alcohol) and health assets (leisure facilities, green spaces).

Compute

- University of Liverpool has invested in HPC capability, and also has collaborations with the **Hartree Centre**. Current capabilities include: (i) BlueGene/Q, which (131,072 cores), and is currently being replaced with even greater computing capacity; (ii) Blue Wonder (11,520 cores); and (iii) 8PB storage; and additional platforms for big data analytics.

Translational and innovation capabilities

Links with industry

- Also, located within Liverpool is **AIMES**, a commercial data centre service provider, which provides N3 secure data storage at Wavertree Technology Campus. In addition, within Connected Health Cities, AIMES hold identifiable data in one secure set of servers, and has created a discrete and ‘ring-fenced’ store for anonymised data with secure on-site computers hosting the relevant statistical software. Aggregated outputs can be generated and ‘published’ from this system, with appropriate controls and permissions.

Links with NHS

- University of Liverpool works closely with **NHS Digital** and have obtained approval for access to Liverpool CCG data stored in a scalable system developed by AIMEs (work in progress).
- University of Liverpool partnerships with local NHS Trusts, four of which (Alder Hey Children’s Hospital, the Royal Liverpool and Broadgreen University Hospitals NHS Trust, Mersey Care NHS Foundation Trust and Wirral University Teaching Hospital Foundation Trust) have been awarded **NHS England Global Digital Exemplar** status, will provide opportunities to align their joint applied research programmes and answer important clinical and public health questions using new informatics and technological methods.
- Innovative approaches introduced include digital dictation and transcription for electronic patient reported outcomes (RLUH), ‘Person Held Record’ test system with NHS England, NHS Digital and the Government Digital Service (RLUH), HealthIntent platform combining hospital and GP data into the shared health record for chronic diseases including asthma, diabetes and COPD (WUTH), structured clinical terminology based data capture (AHCH, LHCH), improving the patient experience using IBM Watson technology (AHCH), co-production of digital technology services (Mersey Care), collaboration with Stanford to develop a suicide prevention app to help reduce death by self-harm (Mersey Care), NHS England national PROMs pilot (LHCH).
- Through **Liverpool Health Partners**, the University links to six additional NHS Trusts and the **NIHR CLAHRC North West Coast** (*Director: Gabbay*) cements further relationships with eight local authorities, six CCGs and 17 NHS Trusts.

Opportunities

- University of Liverpool is developing knowledge, guidance and platforms for efficient trials that can be shared across research groups/programmes.

Potential gaps

- Currently, the University of Liverpool has not got any significant research activities in citizen-driven/digital health but it has plans to expand on this area.

Other relevant capabilities

Research themes

Discovery Science

- **Infectious Disease** is one of three University of Liverpool’s research themes recognised for its research excellence. An interdisciplinary approach brings together medical and veterinary science to focus on diverse challenges like new medicines for children and tackling HIV disease, through to developing diagnostics, treatments and vaccines for both humans and animals. Major programmes in health and biomedical informatics include:
 1. the **“Integrate”** programme (*O’Brien; Wellcome Trust/DH-funded*) to accelerate recognition, investigation and control of gastroenteritis outbreaks within the NHS by analysing together real- time regional and national syndromic data from human and animal surveillance systems in combination with rapid molecular diagnostics and next-generation sequence bioinformatics data on pathogens causing diarrhoea;
 2. the **Small Animal Veterinary Surveillance Network** (*Radford*), a unique research and syndromic surveillance system for companion animals established in collaboration with the British Small Animal Veterinary Association and essential for understanding national human risk for zoonotic pathogens, including antimicrobial resistance;

3. **vaccines** programmes (*French*) aimed at developing efficient methods for evaluating new vaccines; and
 4. **methods** development to model **spread** and **evolution of infectious diseases** in collaboration with the Liverpool Centre for Mathematics in Healthcare (*Sharkey, O'Brien*).
- The **Centre for Genomic Research** (*Hertz-Fowler*) includes a dedicated team of 15 experienced bioinformaticians and software engineers working in conjunction with laboratory specialists ensuring that large and complex genomic data sets can be analysed efficiently and effectively.

Precision medicine

- The **Wolfson Centre for Personalised Medicine** and the **MRC Centre for Drug Safety Science** combine expertise in clinical pharmacology with 'omics technologies to identify biomarkers which allow for: disease stratification; identification of responders and non-responders to medicines; and development of algorithms (combining clinical and genomic information, together with PK/PD and PBPK modelling) to individualise dosing (precision dosing). This has already led to the development of genotype-guided dosing for warfarin (being implemented in primary care), the use of HLA-genotyping to prevent serious adverse drug reactions (e.g. HLA-B*57:01 in abacavir hypersensitivity) and the development of multi-biomarker protein and miRNA panels which can identify patients at risk of serious drug-induced liver injury which outperform conventional transaminase measurements.
- *Professor Pirmohamed* formed and leads the **UK Pharmacogenetics and Stratified Medicine Network**.

Population and Public health

- University of Liverpool hosts the **MRC North West Hub for Trials Methodology Research** (*Director: Williamson*), which undertakes evaluation of approaches to e-recruitment and data sharing; whilst the **Clinical Trials Research Centre** is delivering a number of **NIHR-funded e-trials** including ISDR, CF START, REACT and OVERT, which are using informatics in a way similar to the GSK Salford Lung Study; and developing web- and app-based solutions to collect patient reported outcome data.
- For example, the NIHR-funded ISDR study uses informatics similarly to the GSK Salford Lung Study, randomising 4,500 patients from the cohort of 25,000 individuals in the NHS diabetic eye screening programme within Liverpool. Data are predominantly obtained from routine care, both primary and secondary care. CF START is an NIHR-funded trial, delivered through the UK CF Registry, recruiting through all CF Centres using data currently collected by the registry. The NIHR-funded REACT trial recruits, consents, collects all data and provides the intervention entirely online. The NIHR-funded OVERT trial makes use of an eDiary (available over the web, with an "app" in development) to collect patient reported outcome data.

Citizen-driven/Digital health

- University of Liverpool researchers are developing deep learning methodologies using EPRs to develop personalised renal function monitoring protocols, which will be linked to non-invasive renal function assessment using sensor technology and feedback of results to patients and their clinicians using mHealth platforms.

Learning health systems

- The **Department of Biostatistics** hosts the University's Connected Health Cities **Healthcare Data Laboratory** (*Bodger*), linking its critical mass of analytical expertise with a programme of engagement of front-line clinicians across the NW Coast, supporting co-production of clinically-validated analytics and data visualizations for defined patient pathways.

- University of Liverpool researchers have mapped ‘hot spots’ of emergency admissions for alcohol-related conditions at small area level, overlaying aggregated data on benefits, crime, housing, markers of deprivation, location of GP surgeries and off license premises. Building on this concept of small area mapping to promote local engagement, knowledge of an individual’s place of residence could facilitate improved access to specific services, e.g. physical location of a community clinic, allowing healthcare teams and others to individualise care plans.
- In the “**Integrate**” programme, University of Liverpool researchers analyse on a daily basis national syndromic surveillance data from NHS111 in conjunction with data from SAVSNET and laboratory results to detect geospatial anomalies that might indicate the emergence of an outbreak and deliver results every morning to the Consultants in Health Protection on the front-line, in time for them to respond quickly.

Disease areas

- University of Liverpool has expertise in disease areas that include: infectious diseases, child health, musculoskeletal, CVD and gastrointestinal.

Analytics

- The **Departments of Biostatistics** is undertaking research in areas such as joint modelling of longitudinal and time-to-event data, multivariate data analysis, multi-source evidence synthesis, and statistical genetics and pharmacogenetics. Combined with a strong group of geographers with expertise in spatial statistical analysis (*Singleton, Lloyd*) and a cross-Faculty **Analytics Special Interest Group** (*Maskell*), this offers analytical support and innovative methods development.
- University of Liverpool researchers are developing analytics to identify and improve the **health inequalities** impact of public policies. For example, development of alternative rigorous epidemiological methods to the classical RCT to demonstrate the health inequalities effects of public policy decisions. This includes applying innovative modelling and simulation techniques (*O’Flaherty*) to investigate the impact of structural policies to improve diets on disease burden and equity. Also, they are applying modern methods for causal mediation analysis in longitudinal data tracing social pathways to and from health inequalities in children and identifying effective interventions and policies to reduce health inequalities (*Taylor-Robinson*).

e-Infrastructure

- University of Liverpool’s facilities consist of an HPC cluster with 2,000 cores of freely accessible compute cores and a Condor pool with approximately 1,400 cores to support analytical workflows involving large numbers of independent and relatively short runs. The cluster includes a large memory server with 2TB of memory. All of the cluster’s servers are connected via a high-performance interconnect and have shared access to approximately 200TB of storage. Plans are underway to double the capacity of this cluster by mid-2017 with a further doubling in capacity scheduled for the fourth quarter of 2018 once a planned new data centre is completed.
- Separate to the centrally managed HPC, the Institute for Integrative Biology recently purchased servers and storage for general research support. This includes twelve high-specification compute nodes: ten with 64 cores, 512GB RAM and 2TB SSD-based scratch space; and two with 96 cores, 1TB RAM and 2TB SSD-based scratch space. There is also 1.8PB storage capacity accessed by all compute nodes, connected via 10Gb Ethernet, with 40Gb Ethernet bridging to off-site backup storage.

Training

- University of Liverpool offers **MSc programmes** in: Computer Science; Geographic Data Science; Big Data Management; and Public Health. They are also in the process of developing a bespoke MSc in Health and Biomedical Informatics.
- Integrated **PhD in Data Analytics** delivered via the **ESRC CDT in Data Analytics and Society Centre** directed by Leeds with the Universities of Liverpool, Manchester and Sheffield.
- It also works in collaboration with local NHS Trusts to provide **CPD courses**.

Public Involvement/Engagement

- *Professor Gamble* has led NIHR/INVOLVE-funded research around public engagement and involvement.
- The Patients' Perspectives theme (*Lead: Young*) of the MRC Trials Methodology Research Hub includes a portfolio of research related to health informatics issues, e.g. e-recruitment, consent, retention and outcome measurement, to identify patient-centric solutions.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@HeRC (led and based in the University of Manchester) – award partners include the Universities of: Bradford, Lancaster, Liverpool, Newcastle, Sheffield and York.
- N8 Research Partnership (Universities of Durham, Lancaster, Leeds, Liverpool, Manchester, Newcastle, Sheffield, and York)
- ESRC Consumer Data Research Centre
- London School of Hygiene and Tropical Medicine
- University of Oxford
- Hartree Centre at Daresbury

NHS and other government organisations

- Royal Liverpool University Hospital
- Wirral University Teaching Hospital
- Alder Hey Children's Hospital
- Liverpool Clinical Commissioning Group
- Mersey Care NHS Foundation Trust

Data providers

- CPRD
- NHS Digital
- Public Health England

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:

Industry

- Pharma: Roche, Novartis, UCB and Eli Lilly
- Tech: AIMES, SAP

International

- COMET Initiative: The University leads this consortium comprising the Universities of Oxford, QUB, Bristol and the Centre for Medical Technology in Practice in Baltimore. COMET is a collaboration with many organisations, including EMA, EUPATI, EFPIA, EUnetHTA, HTAi, OECD and CDISC.
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR

University of Manchester

Top-level insights

- The University of Manchester has a strong history in translating computer science and software engineering into eHealth innovations. Its Sol builds on the Farr@HeRC research activities. More specifically, they are proposing to: consolidate the informatics for understanding phenotypes, populations and systems, including the research into learning health systems; and integrate the whole spectrum of informatics in targeted research.
- The University of Manchester has recently merged its health and biomedical faculty to establish the Faculty of Biology, Medicine and Health (FBMH), which has the formation of an integrative informatics and data-sciences research strategy for health at the top of its agenda.
- Informatics education has a strong tradition at the University of Manchester that leads prominent UK health and biomedical informatics education and professionalisation programmes.
- The University of Manchester sits at the heart of radical changes in UK health and social care delivery: the devolution to Greater Manchester (GM) of control over NHS budgets and data governance; and the Greater Manchester Connected Health Cities pilot of learning health systems delivered by the NHTA.
- Farr@HeRC have a very active PPI/E programme with a dedicated team of professionals and a public engagement campaign, #datasaveslives, designed to highlight the positive impact of health informatics research.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The University of Manchester has a strong track record and substantial investments in biomedical informatics that includes the **Farr@HeRC**, which currently focuses on health informatics at the level of populations and systems, for: discovery science with electronic health records; and interventional informatics, particularly learning health systems methodology.
- Farr@HeRC has helped to develop the **Connected Health Cities** (£20M DH) model, which the NHTA is piloting across four regions for integrating and analysing multiple sources of health data consistently. This connectedness is designed to support scientific research and local service development (e.g. risk stratification for targeting services) in synergy.
- It has a world-renowned computer science with close links to a breadth of biomedicine. This has given rise to world-firsts in computable medical terminologies, medical imaging science, and bioinformatics (methods, software and education).

Knowledge/Expertise

- The University of Manchester has strong expertise in a very wide range of methodological areas from biostatistical and machine learning methods for endotype discovery for stratified medicine to real world clinical trials digital health and learning health systems.
- Informatics education has a strong tradition at the University of Manchester, which offered the first MSc in bioinformatics and the first dual bio-health informatics courses – with >5,000 students in the past 10 years. The University of Manchester investigators lead prominent UK informatics education and professionalisation programmes, including the Farr doctoral training network, ELIXIR and NHS Modernising Scientific Careers Clinical Bioinformatics.

Data

- The University of Manchester investigators have access to five UK-based birth cohorts focusing on asthma (Avon Longitudinal Study of Parents and Children, Ashford and Isle of Wight cohorts, Manchester Asthma and Allergy Study and the Aberdeen Study of Eczema and Asthma To Observe the Effects of Nutrition) through the **MRC STELAR Consortium**. Individually, each cohort has great depth in terms of exceptionally well-characterised subjects, a wealth of environmental exposure measures and genetic and other biological data. Data are available on >14,000 children with many repeat measures over time. This cohort has recently expanded to a total of 17 UK/US birth cohorts in the \$100m NIH ‘Children’s Respiratory Research and Environment Workgroup’ (CREW) programme (Buchan).
- Access to data generated through the Manchester-based MRC Stratified Medicine Consortia (PSORT, MASTERPLANS and MATURA – see *section below*) is also available.
- Finally, data that will become available through the Health North programme, which is operating across North England (15M population).

Compute

- The University of Manchester has access to N8 HPC Tier 2 facility.

Translational and innovation capabilities

Links with industry

- Through the cover 50 companies are involved in the associated “connected health ecosystems”.
- A number of ICT companies including Cisco, Siemens and BT are partners with the University of Manchester in a £16M Internet of Things demonstrator (**CityVerve**, £16M IUK).

Links with NHS

NIHR Biomedical Research Centre

- The **NIHR Manchester BRC** (£28.5M, 2017) is a partnership between the University of Manchester and the Central Manchester University Hospitals NHS Foundation Trust. Its areas of strategic priority include mental health, dementia and cardiovascular disease. It also has a focus on cancer, hearing, skin, respiratory, and musculoskeletal diseases. The Manchester BRC does not seem to have developed significant informatics research activities.

Academic Health Sciences Centre

- The **Manchester Academic Health Science Centre (AHSC)** (*Director: Ian Greer*) is the only AHSC outside the SE and is a partnership between the University of Manchester and:
 - the University Hospital of South Manchester NHS Foundation Trust (UHSM)
 - the Christie NHS Foundation Trust
 - the Manchester Mental Health and Social Care Trust
 - the Salford Clinical Commissioning Group
 - the Salford Royal NHS Foundation Trust (SRFT)

The MAHSC encompasses the first fully digitally enabled NHS Trust in England (SRFT), with >500k patient records across primary and secondary care. Using the Salford EHR, trial outcomes and clinical safety are measured in real-time, opening a new global market for clinical trials. For example, the **Salford Lung Study** (£36M GSK investment) is the largest commercial trial in the UK and is unique being the first large, prospective, real-world trial conducted with a pre-license medicine, across a large population in a single geographic setting.

The MAHSC Director has led the creation of the Northern Health Science Alliance (NHS) bringing together the eight largest research active leading universities and NHS Trusts in the North of England. MAHSC is committed to working with the NHS in areas of complementary strength and priority across the North such as pharmacology, ageing research and health inequalities, providing an immediate forum for collaboration on a wider geographical footprint.

Academic Health Sciences Network

- SRFT is also home to NHS Quest and hosts the **Greater Manchester Academic Health Science Network** (GM AHSN).
- MAHSC and the GM AHSN have entered a partnership with Hitachi to combine the informatics know-how with Hitachi's expertise in building platforms for innovative applications. This partnership offers huge potential for the creation of new solutions for the healthcare system and requires the combination of innovation and implementation that an AHSC and AHSN can jointly provide.

NHS 100,00 genomes project | Genomics England Ltd

- The **Greater Manchester NHS Genomics Medicine Centre** is designated for both cancer and rare diseases and is led by Central Manchester University Hospitals NHS Foundation Trust.

Other

- The University of Manchester is part of one of the world's first civic, place-based demonstrator of a learning health system, the **Greater Manchester Connected Health Cities** pilot, which brings together the Manchester AHSC, AHSN, NHS and local government.
- The **Spin-in Lab** is a new innovation environment to build tri-partite partnerships between industry, NHS, and the UoM to accelerate the development of digital health technologies. This model enables companies to develop, test and validate digital health products and services using real-world health data and access to domain specific data analytics capabilities and expert advice on the design and development of their ideas. The most recent Spin-in Lab project is a Farr@HeRC collaboration between **Intelligent Medical Objects** (IMO), Salford Royal NHS Foundation Trust and the University of Manchester to support a research and innovation project to advance the automation of coding diagnoses from digital clinical text.
- The University of Manchester also has a partnership with **NICE** that is developing further around digital futures of trustworthy, actionable evidence, including app validation.

Potential gaps

- None identified.

Opportunities

- The University of Manchester has the capability to advance precision medicine by aligning its strengths in methodology across endotype discovery, trials and citizen driven health. For example, enhancing endotype discovery with deeper analytics for better stratification; and engineering informatics interventions to support better informed, patient-driven personalisation of healthcare interventions.
- It could provide unique expertise and disseminate good practice in research methodology of learning health systems and its experience through the Connected Health Cities programme.
- DataWell is a new health information exchange being rolled out across Greater Manchester, delivering integrated records across all care organisations for 2.8m citizens. DataWell connect to the Farr@HeRC TRE for research and has clinical study feasibility, recruitment and randomisation services (FARSITE). Across North England, via the Connected Health Cities Hub at Manchester, other Datawell-like health information exchanges are working toward large-scale (15m population) data exchange.

Other relevant capabilities

Research themes

Discovery Science

- **IUK Discovery Medicines Catapult** (£50M) – a new, national centre of applied R&D expertise to promote and support innovative, fast-to-patient drug discovery in the UK through collaborative projects across the community.
- **Stoller Biomarker Discovery Centre** (*Whetton*, MRC £16M): aims to identify biomarkers that could be used to diagnose or inform treatment of diseases such as cancer and arthritis. The Centre builds on research already carried out in the Manchester Cancer Research Centre, including discovering new markers for the earlier detection of ovarian cancer – crucial in starting early treatment to save lives.

Precision Medicine

The University of Manchester has three MRC Stratified Medicine awards (total £14.1M). These are:

- **PSORT – Psoriasis Stratification to Optimise Relevant Therapy** (£5M; *Professor Chris Griffiths*): aims to develop tests that can be used in the clinic to help direct personalised treatments and in particular, biologics that work by targeting specific parts of the immune system.
- **MASTERPLANS** (£4.1M; *Professor Ian Bruce*): aims to identify factors /biological predictors that predict which patients respond well on any particular lupus treatment.
- **MATURA – Maximising Therapeutic Utility for Rheumatoid Arthritis** (£5M; *Professors Anne Barton and Costantino Pitzalis*, UoM and QMUL): aims to search for biological and genetic markers in blood and joints which could be used as clues to predict how patients will respond to disease-modifying drugs including methotrexate and anti-TNF biologic drugs. If successful it is estimated that a stratified treatment approach for this condition could save the NHS £13-18M a year.
- **Manchester Molecular Pathology Innovation Centre** (£2.9M; MRC-EPSRC) is developing biomarker based molecular pathology tests will be a major focus of the Manchester node with the initial work aimed at creating tests to diagnose, pick the right treatment and asses the response to treatment for a range of inflammatory conditions including rheumatoid arthritis, lupus and psoriasis.

Citizen-driven/Digital health

- Through the MRC-supported **ClinTouch** and **CareLoop** award (*Lewis, Ainsworth* MRC £1M), UoM investigators have developed a mobile mental health symptom monitoring system which now links patient self-reported observations to both clinical trials and NHS workflow systems.
- The UoM **Wearable Clinic** (*Peek*, EPSRC £1.6M) aims to develop a set of software tools for smartphones and tablets. This will help patients with long term conditions, together with their carers and doctors, to better manage their health in daily life, respond more quickly to changes in symptoms and prevent fall back episodes.

Public and Population Health

- MRC **STELAR consortium** (*Buchan, Custovic – now at Imperial College*) is supporting multi-site, multi-dataset (birth cohort combined with molecular genetics data) epidemiologically oriented health informatics research and statistical machine learning, to capitalise on the rich resource available within the cohorts.
- The University of Manchester has expertise in pragmatic trials (*van Staa*) and have received funding from GSK (sponsor of Salford Lung Study) for the conduct of an observational study comparing the generalisability of SLS patients to those in CPRD.

Learning Health Systems

- The **Greater Manchester Connected Health Cities** pilot of learning health systems is one of the world's first civic, place-based demonstrators of a learning health system for North England's 15m population. To deliver this, University of Manchester has come together with local government (GM Combined Authority) and the GM NHS to form a civic partnership known as: Health Innovation Manchester. This partnership under local government devolution arrangements, merges the Manchester AHSN and AHSC, and integrates NHS and other civic data sources in a data sharing authority (GM Connect) and system (Datawell software system). All partners are working in a novel chain of civic analytics operating under new local by-laws of a "duty to share data" between organisations for citizen benefit.

Disease areas

- University of Manchester's strategic priority disease areas are: cancer, chronic inflammatory diseases (musculoskeletal, respiratory, dermatology), cardiovascular, dementia and mental health.

Analytics

- The **Asthma e-Lab** is a secure web-based research environment to support consistent recording, description and sharing of data, computational/statistical methods and emerging findings across the five MRC STELAR UK birth cohorts. The e-Lab serves as a data repository for its unified dataset and provides the computational resources and a scientific social network to support collaborative research. All activities are transparent, and emerging findings are shared via the e-Lab, linked to explanations of analytical methods, thus enabling knowledge transfer.
- Reaching further into Computer Science and Mathematics University of Manchester (*Peek, Moulton and Taylor*) are driving EPSRC's **UK Health Data Analytics Network**. This brings together a broad community of over 400 health data scientists across the UK that will work together to map the UK health data analytics landscape and to frame the national strategy for health data analytics research.
- EPSRC **HealTex** (£0.4M) clinical text mining network. Its aim is to build a UK-wide multi-disciplinary research network in order to explore the barriers to effectively utilising healthcare narrative text data, road-map research efforts and principles for sharing text data and text analytics methods between academia, NHS and industry.

e-Infrastructure

Compute

- The University of Manchester has access to **N8 HPC Tier 2 facility** (SGI HPC cluster with a total of 332 compute nodes – 316 nodes (5,056 cores) with 4 GByte of RAM per core (each node has 8 DDR3 DIMMS each of 8 GByte ie 64 GBytes of memory per node and 6 nodes (256 cores) with 16 GByte of RAM per core; 174 TBytes Lustre v2 parallel file system for storage).

Trusted Research Environment

- Sensitive personal data are handled by **Farr@HeRC TRE**, an OpenStack private cloud across 480 cores, 4TB of RAM and 180TB of storage with 40Gb/s interconnect, by provision of tailored virtual machines. The TRE is certified to IG Toolkit level 2 and will undergo ISO27001 preliminary audit in Jun 2017 with full audit in Nov 2017. It uses Arkivum for archive services (48Tb available). Remote access to TRE uses two-factor authentication (RSA SecurID), and there is a physically secure onsite room for accessing the most sensitive data.
- The Farr@HeRC TRE is now integrated into the **Greater Manchester Datawell**, an integrated electronic health record for primary, secondary and tertiary care, covering 2.8M population. The Farr Institute HeRC TRE provides the data and analytics infrastructure for the Greater Manchester Connected Health City and plays a key part in data analytics for Greater Manchester health and social care devolution. The Farr@HeRC TRE is now an active component of the health system it is connected to, acting as a hub through which data can be exchanged between NHS sites, research infrastructure and patients. Across the North of England, work is ongoing to network the data analytics centres of the four Connected Health Cities with the Farr@HeRC TRE, creating a federation of population health centres across the North of England.
- Other classes of data can be handled by the University of Manchester Compute Shared Facility (9288 core 250 TB fast/scratch), Data Processing Shared Facility (>800 cores; most nodes 512 GB RAM, 700 TB fast scratch), both connected to Research Data Storage platform (3PB), connection via a dedicated, secure, fast (20 Gb/s) network.

Training

The Farr@HeRC has a wide range of training programmes and activities including:

- **Capacity-building programme** and purpose-built training suite used to deliver Farr courses UK-wide;
- **MSc in Health Data Science**, one of 3 in the UK, attracts UK and international students;
- **Doctoral Training Network** has attracted 31 students: 2 from the University of Manchester have won the British Computer Society Early Career Award;
- Integrated **PhD in Data Analytics** delivered via the highly multidisciplinary **ESRC CDT in Data Analytics and Society Centre** directed by Leeds with the Universities of Liverpool, Manchester and Sheffield;
- **PhD programme in Quantitative and Biophysical Biology**.
- **CPD courses** in analytics with electronic health record data have been delivered to 270 students;
- **Executive- level short programme** for current and aspiring leaders across sectors (e.g. Chief Clinical Information Officers); and
- **Workforce development** Connected Health Cities and the scoping work for the new **NHS Digital Academy** (*Buchan on Steering Group*).
- In social science, the University of Manchester (Elliot) drives UK-wide training in the **SURE consortium** with the Farr Institute and a new ESRC doctoral training network across North England in new forms of data.

Public Involvement/Engagement

- The Farr@HeRC has a **PPIE team** of professionals to make sure that people have the opportunity to comment on their research.
- They been the masterminds behind the **#datasaveslives**, a public engagement campaign designed to highlight the positive impact of health informatics research on public health. The campaign has gained support from Ministers and politicians (George Freeman MP, Shona Robison MSP, John Barron MP), the public, policy leaders (Bedirhan Ustun, WHO) and researchers alike. The campaign has been picked up by other organisation such as CRUK, PHE and the government backed #empower campaign for health data use in research. #datasaveslives has also been adopted by the Health North pilot CHC as their leading tool for public engagement.
- The Farr@HeRC has also an active patient and **public involvement forum (H@PPI)**. The forum meets at least quarterly and functions include reviewing grant proposals, identify patient priorities for research, setting standards for PPI in HeRC projects and monitoring the impact of PPI.
- The University of Manchester is carrying out research to explore important question about the use health records in research; and to what extent should patients control access to their health records. To help with this research, it has convened two cross-sections of 18 adults from Greater Manchester, **“citizens’ juries”**, to hear from and ask questions of experts, and to deliberate on these issues and validate the outcomes of the process. The project is funded by the Farr@HeRC and the NIHR Greater Manchester Primary Care Patient Safety Translational Research Centre, and supported by the Information Commissioner’s Office. A report summarising the design and findings of the two citizens’ juries has been published and presented at a post-jury workshop.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@HeRC (led and based in the University of Manchester) – award partners include the Universities of: Bradford, Lancaster, Newcastle, Sheffield and York.
- Farr@London (UCL), Farr@Scotland and Farr@CIPHER(Swansea)
- N8 Research Partnership (Universities of Durham, Lancaster, Leeds, Liverpool, Manchester, Newcastle, Sheffield, and York)
- MRC Stratified Medicine Consortia – MATURA; PSORT; MASTERPLANS
- ESRC CDT – new forms of data with Leeds, Liverpool and Sheffield
- EPSRC Wearable Clinic
- EPSRC UK Health Data Analytics Network and HealTex

DH and NHS

- NIHR Manchester BRC
- Manchester AHSC and AHSN – the only AHSC outside the south east
- Health North: Connected Health Cities – over 50 companies are associated with this initiative
- NICE

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:

Data providers

- Public Health England
- NHS Digital
- CPRD

Industry

- Pharma companies: GSK
- Technology companies: Cisco, Siemens, BT, Cerner, IMO

International

- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR
- Collaboration with NIH that has expanded the MRC STELAR consortium from 5UK to 17 UK/US birth cohorts
- US BD2K initiative – MRC STELAR Asthma eLab Research Object design adoption

Newcastle University

Top-level insights

- The Newcastle University has substantial strengths and unique investments in digital systems as evidenced by the hosting of the National Innovation Centre for Data and the establishment of Newcastle University's Centre for Health and Bioinformatics. This latest investment has enabled the identification and coordination of local expertise, the integration of research teams with common aims, enhanced support through platforms such as the Bioinformatics Support Unit and targeted training for the next generation of informaticians working in the health sciences.
- Other strengths include expertise in human-computer interaction; cloud computing and cyber-security evidenced by the funding awards including: the Next Stage Digital Economy Research Centre; Centre for Doctoral Training in Digital Civics; and the GCHQ Academic Centre of Excellence.
- It has also been leading efforts on data access and governance through initiatives aiming to develop an open-source analysis tool for sensitive data or data under strict governance controls.
- There are strong industry collaborations, in particular with the technology firms including IBM, Microsoft and Facebook.
- Newcastle University has been working closely with the Farr@HeRC (Manchester). Through Connecting Health Cities and the Great North Care Record, it aims to establish the platform for a North East and North Cumbria Learning Health System that will enable improvements in health and social care for the citizens of the region.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

Newcastle University's major inter-disciplinary investments in the areas of biomedical research and population health with relevance to informatics include:

- The **Digital Institute** (*Director: Professor Paul Watson*) works with researchers across all disciplines at NU to transform their research through the application of advanced digital technologies. It has collaborated with many healthcare projects, including: activity recognition through machine learning for cardiovascular disease; wearables for diabetes monitoring; computer games in stroke therapy; and cloud computing for genomics. The Digital Institute also hosts **EPSRC's Doctoral Training in Big Data and Cloud Computing** (*see NU training section below*).
- In 2016, the Newcastle University invested £30M to create the **National Innovation Centre for Data (NICD)**. NICD brings together industry, the public sector and world-leading academics to develop the skills, ideas and resources needed to exploit the opportunities offered by the explosion in digital data.
- The **Centre for Health and Bioinformatics (CHaBi)** is a cross-faculty research centre that brings together researchers in bioinformatics and health informatics and was set up to address: the identification and coordination of local expertise; the integration of research teams across the University with common aims towards collaboration; training the next generation of bioinformaticians and health informaticians. Based within CHaBi is the **Interdisciplinary Computing and Complex Systems group (ICOS)** (*Wipat,*

Krasnogor), which undertakes research at the interface of computing science and complex biological and medical systems. The group has a track record in software systems development for biological and medical applications, with over twenty tools currently provided for use.

- Newcastle University has hosted a **GCHQ Academic Centre of Excellence (ACE)** (*van Moorssel*) in **Cyber Security Research** since 2012. The ACE counts 25 graduated PhD students since its inception and currently comprises ~40 researchers. Grants include: £1.1M European Research Council Fellowship (*Feng Hao*); £4M GCHQ/EPSRC Research Institute in Science of Security; and £2.5M NCA/Mets Police/EPSRC Cloud Crime Centre. Newcastle ACE is the founding member of **CyberNorth**, an initiative to improve cyber security practices in the North of England and to create entrepreneurial and other activities in cyber security that benefits the economy of the region.
- **Open Lab** (*Olivier*) is human-computer interaction, social and ubiquitous computing research group in **the School of Computing Science**. It comprises six academics, 35 post-doctoral researchers and 70 doctoral trainees spanning fields from computer science and engineering, to design, psychology, sociology, education and clinical sciences. Open Lab has a portfolio of externally funded biomedical research ranging from wearable computing for self-management of chronic illness, new media platform for collation and sharing of citizen health experiences, to design methods for clinical intervention for non-standard users and the design and deployment digital tools for clinical research. One such tool is the *Open Movement AX3* physical activity sensor, distributed by NU spin-out Axivity Ltd, and recently used in UK Biobank's collection of 7-day physical activity in 100K participants.

Knowledge/Expertise

- Newcastle University leverages interdisciplinary expertise in computational intelligence, networks, knowledge management, and data integration to problems in biomedical and neuro-informatics.
- Also, particular expertise is on the analysis of big data, including both static data (e.g. from genomics) and streaming data (e.g. from medical sensors wearable devices). It is one of the pioneers in the use of scientific cloud computing to speed-up data analysis.
- Activities on data access and governance are through initiatives such as:
 - **DataSHIELD** (also supported by the Farr Institute) (Burton and Murtagh involved in its initial development) – an open-source analysis tool for sensitive data or data under strict governance controls;
 - **BioSHaRE-eu** – funded initial technical development of DataSHIELD and formal exploration of its interface with the complex governance frameworks (including citizen engagement) and societal structures it operates within;
 - **METADAC** (*Professor Murtagh* chairs and is PI of METADAC (MRC, ESRC, Wellcome)) – is the first national multi-cohort Data Access Committee (BBMRI- LPC).;
 - **58FORWARDS** – operating via METADAC to develop and maintain underpinning technical systems; and
 - **BBMRI- LPC** – on the evaluation of data access and open science strategies for Biobanks at the European level. Professor Murtagh leads evaluation of “data access and open science strategies for Biobanks at the European level”.
 - **RD-Connect** – links genomic data with registries, biobanks, and bioinformatics tools to produce a central resource for rare diseases.

- Newcastle University has a long track record and excellent expertise in ageing research, which is co-ordinated via the cross faculty, inter-disciplinary **Institute for Ageing**, Europe's largest such centre with over 700 academics, clinicians and researchers working on all aspects of ageing research and innovation from medical, biological and cellular, to psychosocial, economic and environmental. Building on this, they also received a government award in 2015 to establish a **National Innovation Centre for Ageing (NICA)**.

Data

- The **MRC-Wellcome Trust Human Developmental Biology Resource (HDBR)** is organised from two sites: the Institute of Genetic Medicine, Newcastle, and the Institute of Child Health, London. The HDBR is an ongoing collection of human embryonic and fetal material ranging from 3 to 20 weeks of development. Tissue, slides, RNA/DNA and tissue arrays are available to the international scientific community. Gene expression data emerging from HDBR material is added to our gene expression database which is accessible via our HUDSEN (Human Developmental Studies Network) website. This amongst others provides an online resource for studying prenatal human brain development.

Compute

- University of Newcastle has two data centres (on and off-site) and has invested in a VNWare virtual infrastructure. Enterprise storage has recently been refreshed and provides over 2PB of usable storage. The University is currently building a new HPC service that will have 3,500 cores and a dedicated staff to support researchers' use of the system.

Translational and innovation capabilities

Links with industry

- Led by industry, the **National Innovation Centre for Data (NICD)** aims to enable businesses to extract value from smart data and increase competitiveness. A key partner will be Dynamo, the industry-led initiative established to grow the technology sector in the North East.
- The **GCHQ Academic Centre of Excellence (ACE) (van Moorsel) in Cyber Security Research** has received funding from: Ineqe, ARM, e-Therapeutics, Barclays Bank, Microsoft, IBM, Facebook and Hewlett-Packard.
- Newcastle University researchers (*Wipat*) lead a joint project with **GSK** to develop integrated datasets for the identification of drug repositioning opportunities

Links with NHS

- Newcastle University hosts the **NIHR Innovation Observatory (Trenell)**, a £10M investment in health innovation horizon scanning. The centre is applying text mining, natural language processing, machine learning and cognitive analytics to data sources to provide a real-time picture of the global health innovation landscape. NIHRiO provides strategic insight for the Department of Health and Social Care, NHS England and NIHR. The centre also reviews all new drug and non-drug therapies to be appraised by the NICE.
- The **North-East Quality Observatory Service (Whitty)** uses analytical skills and epidemiological expertise to provide quality measurement services across the North East and Cumbria. As an example of a highly regarded service to the NHS, it interprets the quarterly hospital mortality data and provides reports for providers. The reports cover all aspects of the statistical monitoring of hospital mortality. Individual Trusts can commission "deep dive" investigations from NEQOS for an area of concern.

- In addition to the investment in ageing research mentioned previously, the recently awarded **NIHR Newcastle BRC** (£16.2M) is a partnership between Newcastle upon Tyne Hospitals NHS Foundation Trust and NU with a focus on ageing and long-term conditions including: dementia, liver disease, musculoskeletal disease, neuromuscular disease, skin and oral disease.
- Following the successful **National Institute for Health Research (NIHR) Diagnostic Evidence Co-operative Newcastle (DEC)**, which ran from 2013-2017, Newcastle University has been chosen to host one of the NIHR Medtech and In vitro diagnostic Co-operatives (MICs) from January 2018. The Newcastle MIC will provide a centre for excellence for the development and evaluation of new effective diagnostic tests.
- There is a close relationship with the **North East and North Cumbria Genomics Medicine Centre** within the **Newcastle upon Tyne Hospitals NHS Foundation Trust**.
- The **Newcastle Academic Health Partnership** brings together the Newcastle University with the Newcastle upon Tyne Hospitals NHS Foundation Trust (acute services) and the Northumberland, Tyne and Wear NHS Foundation Trust (mental health and disability). The Partnership's **Health Informatics Group** is overseeing the integration of biomedical data with tissue banks enabling seamless linking of the two Trust's EPRs with laboratory data. Biobanking of DNA, RNA and plasma/serum, built around routine patient episodes is linked to the integrated informatics platform. This has been achieved by merging University biobanking facilities with recently configured pathology services in both NHS Trusts.
- The **North East and North Cumbria Academic Health Science Network (NENC AHSN)** is dedicated to improving healthcare and supporting wealth creation through partnership working and promotion of innovation across the NHS, academia and industry. Cross-cutting programmes in digital health and data integration are supporting Telemedicine Projects in Prenatal Diagnosis and Rural Health in Cumbria.

Opportunities

- Newcastle University ranks among the top 10 UK Universities (based on REF2014) in computer science and informatics. Its research activities through the NU Digital Institute and expertise in the areas of machine learning, cloud computing, cyber security but also data access and governance of cohort and biobank data can help catalyse HDR UK activities in these areas.

Potential gaps

- There does not seem to have significant activity on the methodological design of innovative clinical trials using EHR.

Other relevant capabilities

Research themes

Discovery Science

- The **Institute of Genetic Medicine** is using developments in genomic technologies to enhance insights into the genetic causes of not only rare but also common diseases.
- The **Centre for Integrated Systems Biology of Ageing (BBSRC)** and the **MRC Centre for Ageing and Vitality**.
- The **D-BOARD** project (*Barcardit, Loughlin, Reynard*) aims to discover biomarkers for early diagnosis from a broad variety of 'omics and imaging data from already existing human cohorts and animal models.
- The **APPROACH** project (*Barcardit, Loughlin*) focuses on computational approaches to identifying new phenotypes for osteoarthritis.

Precision medicine

Newcastle University researchers are involved in three MRC Stratified Medicine Cohorts:

- **UK-PBC** – understanding disease mechanisms and targeting therapies in primary biliary cirrhosis (£4.8M; *Professor David Jones*). The consortium aims to recruit 10,000 patients at sites around the UK to gain a better understanding of response to treatment.
- **RA MAP** – the rheumatoid arthritis consortium (£4.5M; *Professor John Isaacs*); co-led by KCL and NU. RA MAP aims to investigate two aspects of the disease. First, the identification of prognostic factors using historical data and cohorts. Secondly, through the TACERA study (a longitudinal RA cohort), a systems immunology dissection of RA pathogenesis, to identify tractable pathways and targets.
- **PSORT** (£5.1M)(MRC)(*CI Professor Chris Griffiths, PI Professor Nick Reynolds Newcastle*) aims to rationalise and optimise current and future biologic therapies for psoriasis in a cost-effective and minimally-dosed manner. To understand the heterogeneity of psoriasis and patient responses to biologic therapy, PSORT is bringing together longitudinal outcome data and samples on large numbers of patients receiving biologics (via BADBIR Bioresource; >15,500 subjects), prospective recruitment of deeply phenotyped discovery and refinement cohorts, coupled to state-of-the-art clinical and laboratory based investigation and sufficient bio-informatic and statistical power to interrogate the multi-dimensional datasets generated.
- **Newcastle Molecular Pathology Node** (*Krasnogor*) (£2.7M; MRC-EPSC) is developing advanced machine learning tools (through ICOS) to provide high-throughput automatic image analysis of immunohistochemistry images and to gain insight into the biology and function of cells in health and disease.

Citizen-driven/Digital health

- **Open Lab** is home to the **Next Stage Digital Economy Research Centre** (£4M; EPSRC) which aims to design, develop, and evaluate new digitally mediated models of citizen participation that engage communities, the third sector, local government and (crucially) the commercial digital economy in developing the future of local service provision and local democracy.

Learning health systems

- The **North East and North Cumbria Connected Health Cities** (CHC) is hosted by the University of Newcastle and has two major aims: (i) to establish a platform for the sharing of health and social care data for patient care and research; and (ii) to support projects, which use informatics in care pathways.
- The **Learning Healthcare Project** (*Foley*), funded by the Health Foundation, has collected exemplars and insights from around the world, exploring key themes, technologies and methodologies and their applicability in the UK. The project identifies the importance of Data-Knowledge, supported by electronic health records, outcomes measurement and analytics, but also Knowledge-2-Practice, through sociological approaches to the diffusion of technology, behaviour change and complexity theory. These methods compliment traditional approaches to dissemination, resulting in health systems that learn from every patient.

Disease areas

- Areas of particular strength in addition to ageing, mitochondrial disease, neuromuscular disease and rare diseases include: inflammatory bowel disease, inflammatory skin disease, primary biliary cirrhosis, non-alcoholic fatty liver disease, pulmonary hypertension, congenital heart disease, inherited renal disease, cancer and arthropathies.

Analytics

- The neuroinformatics stream (*Kaiser*) develops algorithms and tools to inform the diagnosis and treatment of neurodevelopmental and neurodegenerative disorders; and is also involved in the EPSRC/ Wellcome Trust **CANDO project** aimed at developing an implantable device for optogenetic stimulation in epilepsy patients.
- The **i-sense project** (*Wipat, McNeil, Keegan*) is developing cloud computing based approaches for the development of rapid diagnostic systems for infectious diseases.
- To make it easier for users to exploit the scalability offered by clouds, the experts at the Digital Institute have built an easy-to-use tool that enables the design and execution of analytics pipelines by non-experts. This tool – **e-Science Central** – is used on a wide variety of biomedical and health applications in academic research projects. It is also deployed for production use in industry.

e-Infrastructure

Compute

- Newcastle University has two data centres connected by dual 100Gbps fibres (giving 200Gbps inter-connection) running over diverse routes from separate providers. The primary data centre is off-site and hosts all critical services, including the University HPC Service, Storage and VMWare environment. Key services are mirrored to the secondary (on-site) datacentre, and Disaster Recovery backups are taken and kept in a third location. The University has invested in a VMWare virtual infrastructure and has recently installed an additional 20 hosts. As standard, all servers have a 10Gbps connection. Enterprise storage has recently been refreshed and provides over 2PB of usable storage on SSD, ASA and SATA technologies. The University has a 10GBs connection to JANET, currently running at 40% of capacity.
- Newcastle University has built a new University HPC service (the Rocket HPC service). The system at the centre of this service has ~3,500 cores, and includes 2 extra-large memory nodes with 1.5 TB RAM and 4TB local disk, plus 6 large memory nodes with 0.5TB RAM and 4TB local disk. The service has dedicated staff to support researchers' use of the system.
- The School of Computing Science hosts an HPC cluster currently composed of 12 computing nodes (Intel Xeon E5-2690 v2/v3) with a total of 232 cores and 256GB/512GB of RAM per node. Two of the nodes have high-end GPU cards (Nvidia Tesla K40). The cluster also has a storage node with 36TB of disk space. The whole cluster is wired with high-speed 56GB/s Infiniband network.

Trusted Research Environment

- To support researchers requiring access to clinical data, the NU IT has developed an Information Governance Toolkit framework, which will allow any project to obtain level 2 compliance. This includes a "Safe Haven" for secure data storage and a dedicated IG lead, who provides specialised advice and support. The IT Security Service in NU offers advice on how to implement security controls mandated by research funders and data providers, ISO 27005 conformant information security risk assessments, supply chain due diligence, server and web application vulnerability scanning, and general information security consultancy.

Training

- Newcastle University's **MSc in Bioinformatics** includes themes such as: synthetic biology; computational systems biology; and neuroinformatics. ~35 students register for the degree each year. There is also an MRes in the Faculty of Medical Sciences which includes bioinformatics modules.

- The Digital Institute hosts **EPSRC's Doctoral Training in Big Data and Cloud Computing** (*Watson*), offers around 11 students per year the opportunity to study for a PhD. It will produce multi-disciplinary experts in the mathematics, statistics and computing science of extracting knowledge from big data, with practical experience in exploiting this knowledge to solve problems across a range of application domains.
- The EPSRC-funded **Centre for Doctoral Training in Digital Civics** (£4.7M) provides cross-disciplinary training to students (55 cross-disciplinary PhD studentships between 2014 and 2022). It is focused upon three challenging application domains for digital civics research: local democracy, education, and public health & social care. There will also be two underpinning technology training programmes: human-computer interaction and security, privacy & trust.

Public Involvement/Engagement

- Newcastle University's Policy, Ethics and Life Sciences Research Centre aims to produce world class research focused on the social and ethical debates around the contemporary life sciences. Researchers at the Centre work with a wide range of academic and community-based partners to develop: rigorous and socially engaged analysis; innovative teaching; and public dialogue.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@HeRC (led and based in the University of Manchester) – award partners include the Universities of: Bradford, Lancaster, Newcastle, Sheffield and York.
- N8 Research Partnership (Universities of Durham, Lancaster, Leeds, Liverpool, Manchester, Newcastle, Sheffield, and York)
- MRC Stratified Medicine Consortia
- MRC-EPSRC Newcastle Molecular Pathology Node
- University of Bristol collaboration on ALSPAC cohort
- METADAC
- UK Biobank
- EPSRC Doctoral Training in Big Data and Cloud Computing

NHS and other government organisations

- NIHR Newcastle BRC
- NIHR Innovation Observatory
- NIHR Newcastle Diagnostic Evidence Co-operative
- Genomics England – North East and North Cumbria Genomics Medicine Centre
- Newcastle Academic Health Partnership
- North East and North Cumbria Connected Health Cities
- North East and North Cumbria Academic Health Science Network
- North East Quality Observatory Service (NEQOS)

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:

Data providers

- NHS Digital

Industry

- Pharma: GSK, Thermo Fisher Scientific
- Tech: ARM, Dynamo, Facebook, IBM, Ineqe, Hewlett-Packard, Microsoft

International

- EU BBMRI ((Biobanking and Biomolecular Resources Research Infrastructure)
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR

University of Sheffield

Top-level insights

- The University of Sheffield strengths in health and biomedical informatics build on its investments in the School of Health and Related Research (ScHARR); the Insigneo Institute; the Sheffield Institute for Translational Neuroscience; the MRC Polaris Centre; and The Centre for Genome Translation.
- Areas of strengths include: population health research which takes a broad, multidisciplinary perspective, ranging from public health and acute care to health economics and decision science.
- Expertise from the University of Sheffield is combined with that of their partners at Sheffield Teaching Hospitals NHS Foundation Trust, as exemplified by Sheffield's NIHR Clinical Research Facility and Biomedical Research Centre.
- University of Sheffield has links with the Farr@HeRC(Manchester) and there are plans to further utilise their wider membership of the Northern Health Science Alliance and their partner NHS Trusts.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

The University of Sheffield's major inter-disciplinary investments in the areas of biomedical research and population health with relevance to informatics include:

- **The Insigneo institute of *in silico* Medicine** is a joint initiative between the University of Sheffield and Sheffield Teaching Hospitals NHS Foundation Trust (*see also NHS Links section below*). It performs cutting edge research in areas of fundamental and applied biomedical modelling, imaging and informatics.
- University of Sheffield's population focus takes a broad, multi-disciplinary perspective and is structured under the **School of Health and Related Research** (ScHARR, average annual research income £9.8M), which incorporates a number of major programmes of work, including: the NIHR School for Public Health Research; the NIHR-funded Technology Assessment Group; the DH- funded Economic Evaluation Policy Research Unit.

Knowledge/Expertise

- The ScHARR is one of the largest and most dynamic Schools of health research employing over 300 staff across the health sciences and public health, including: public health practitioners, clinicians, epidemiologists and statisticians, health economists and modellers, trialists and information scientists, psychologists, sociologists, management experts and political scientists.
- Regarding health informatics, the University of Sheffield has expertise in managing, linking and analysing large datasets, with policies and secure systems in place for data handling.
- There are also expertise in the use of biomedical informatics linked to genomics to support medical decision making and patient stratification covering all essential aspects, including data collection, information extraction, multimodal information management, knowledge extraction, and predictive medicine.

Data

- Examples of the range of population-based datasets University of Sheffield holds or has access to include:
 - the **TARN database**, which has been instrumental in influencing the development of the UK Major Trauma Centre network as well as numerous National Institute for Health and Clinical Excellence (NICE) clinical guidelines;
 - the regional **Yorkshire Health Study**;
 - the local **Sheffield total population general practice cohort** with linked social service data;
 - the **MRC-funded Alcohol Policy Modelling and Evaluation** based on analysis of 201M hospital episode records linked to alcohol sales outlets data on over 100,000 outlets, which has been informing public and political debate on alcohol policy; and
 - the **Summary Hospital-level Mortality Indicator (SHMI)**, now implemented nationally to report on mortality at Trust level across NHS England using a standard and transparent methodology.
- Connected to this is the University's **Data Open Access policy**, supported by the **Research Data Catalogue and Repository (ORDA)** which offers publication of individual datasets, providing a DataCite digital object identifier (DOI) unique for each dataset.
- Sheffield is driving an international consortium including Harvard and Massachusetts Institute for Technology (MIT) to develop cloud-based infrastructure for an experimental data repository for sharing and publication. This scalable commons infrastructure provides collaborative access to data across different domains.

Compute

- The University of Sheffield has dedicated HPC resources for data analytics, modelling & simulation. Furthermore, the Insigneo Institute has links to National (e.g. EPCC) and European (e.g. PRACE) e-infrastructure resources.

Translational and innovation capabilities

Links with industry

- University of Sheffield is a member of the **Avicenna Alliance**, a pre-competitive alliance between biomedical industries and academic institutions aimed to promote the use of modelling and simulation in the development regulatory assessment of biomedical products.

Links with NHS

- The **Insigneo institute of *in silico* Medicine** is a joint initiative between the University of Sheffield and Sheffield Teaching Hospitals NHS Foundation Trust, which coordinates 140 academics and clinical consultants sitting in 28 different departments, and conducting research in five broadly distinct research areas all interconnected into a shared vision, that of predictive medicine.
- The new **NIHR Sheffield BRC (£4M)** is a partnership between the Sheffield Teaching Hospitals NHS Foundation Trust and the University of Sheffield. It focuses on translational neuroscience for chronic neurological disorders and has a cross-cutting theme in genome medicine analytics.

Opportunities

- The University of Sheffield's strengths and expertise in public health research and health economics through the School of Health and Related Research and in *in silico* medicine through the Insigneo Institute (e.g. *in silico* Clinical Trials). These could contribute towards HDR UK's activities in: developing innovative informatics systems capable of a seamless data analytics process for population medicine, but also individualised medicine, through the creation of new clinical decision-support systems for diagnosis, prognosis, and treatment planning / stratification; and improving scalability of subject-specific modelling technologies to enable their use also on very large phase III clinical trials, involving thousands.

Potential gaps

- The University of Sheffield has niche expertise in the areas of discovery science, citizen-driven health and learning health systems. It holds huge amounts of routine data on patients and health services. However, there are gaps whereby this data is not centrally housed and therefore the potential to link data together is lost. Sheffield is ambitious to facilitate central warehousing of large datasets and develop a coordinated approach to managing and linking that data in order to provide a more comprehensive picture of the system and patient outcomes following contact in the system.

Other relevant capabilities

Research themes

Discovery Science

- In terms of discovery science, University of Sheffield has strengths in: genomics and bioinformatics (with the new **Centre for Genome Translation**); medical imaging (with the MRC Pulmonary, Lung and Respiratory Imaging Sheffield (POLARIS) Centre).
- Furthermore, the Sheffield **Institute for Translational Neuroscience** (£20M) houses 200 researchers, including a clinical database of over 1,500 patients, a vital resource of human brain-bank material (central nervous system material from > 850 individuals).

Precision medicine

- There isn't any information in the Sol on any specific research programmes in stratified/personalised medicine.

Population and Public health

- The **MRC POLARIS** (Pulmonary, Lung and Respiratory Imaging Sheffield) imaging facility includes state of the art IT facilities for the management and automated analysis of medical imaging data, especially the unique hyperpolarised gas proton MRI that the POLARIS facility generates.
- The **Sheffield Clinical Trials Research Unit** uses secure information systems that allow remote access data capture, including the collection of identifiable and highly sensitive personal information, and uses controlled data analysis facilities (including virtual machines) set up on secure TUOS computing facilities

Citizen-driven/Digital/Digital health

- University of Sheffield has specific skills in areas such as development optimisation and validation of wearable sensors, digital health, mobile health, and personal health forecasting such as real time prediction using subject-specific models informed by wearable sensors, for example in Parkinson's disease, to advise patients during their daily life.
- The Insigneo Institute has a programme on **Personalised Decision-Support for Heart Valve Disease** (H2020, £4M) that could assist clinicians in difficult areas where extra data might assist the decision.

Learning health systems

- Through the **NIHR Yorkshire & Humber Collaboration for Leadership in Applied Health Research and Care (CLAHRC)** and **Connected Health Cities**, University of Sheffield researchers have collected, cleaned and linked data in order to evaluate urgent & emergency care systems across the region, and genome diagnostics pathways.

Disease areas

- Areas of strength include: neurological, cardiovascular, respiratory, and metabolic diseases. Also, public and population health including urgent and emergency care, mental health and ageing.

Analytics

- The Insigneo Institute offers some mature ***in silico* medicine data analytics** services such as:
 - the VirtuHeart service for non-invasive planning of coronary stenosis (funded by EPSRC and Wellcome/Department of Health and Social Care with £2M), or
 - the CT2S service for the prediction of risk of bone fracture in osteopaenic patients (Funded by the FP7 VPHOP project, £10M).

Furthermore, Insigneo has world-class research programmes in multi-scale modelling (funded by EPSRC, £5M).

- The **POLARIS Imaging Informatics Platform** allows secure handling of all medical imaging data collected in experimental medicine studies, and their automated processing according to the image analytics tools developed by the Insigneo institute
- University of Sheffield has clinical and computational expertise to develop complex algorithms to analyse datasets. For example, through the **NIHR Programme Grant on Vascular Services** (£1.8M), researchers have been analysing a hierarchy of vascular operations. This includes multiple staged operations, comorbidities, complications and outcomes with linked costs and patient pathways, to inform future policy decisions.
- Other analytical capabilities include structural equation statistical modelling, Bayesian hierarchical spatial modelling and complex geographic information system analysis – for example, the ESRC-funded project on **Improving Well-being through Urban Nature** (£1M).

e-Infrastructure

Compute

- The University of Sheffield's Tier-3 resources for data analytics, modelling and simulation include a SGI2000 cluster with 1.5TB of memory accessible by a single process (OpenMP); over 4,000 processing cores accessible through the Iceberg and ShARC Infiniband clusters with a 1PB fast data storage system (Lustre parallel file system), and 24 GPU nodes.
- Through consortia agreements and dedicated grants Insigneo has access to N8 HPC consortium, EPCC ARCHER, and the European PRACE e-infrastructure resources.
- The University of Sheffield's Corporate Information and Computing Services maintain a central storage estate (NetApp) which currently offers 2.4 PB of redundant storage per research group to support project-level sharing of data during the research process.

Trusted Research Environments

- The Scientific Computing group at Sheffield Teaching Hospitals NHS Foundation Trust manages the **Dedicated Research Data Warehouse**, constructed for ethically approved secondary use of clinical data, and the **Integrated Research Management System** which handles governance requirements, patients' allocation, and researchers' allocation for all clinical trials.

The warehouse is derived from the principles of the i2b2 system, but has been completely re-engineered to handle the volumes of data required, currently almost 200M records. The technology is now based on the Apache Spark and Drill software stack with custom interfaces for querying the data. All this data is also replicated into an ElasticSearch server which pre-processes it to support natural language queries on the free text components of the data, typically referral correspondents and radiology reports. These services run in the hospitals data centre and are hosted in its VMWare environment.

- The **POLARIS Imaging Informatics Platform** allows secure handling of all medical imaging data collected in experimental medicine studies, and their automated processing according to the image analytics tools developed by the Insigneo institute.
- Though their membership in the NHTA, University of Sheffield has access to the **Connected Health Cities (CHC)** infrastructure, which includes shared facilities for distributed health data analytics.

Training

- University of Sheffield co-ordinates one of the longest-running **MSc Health Informatics** programmes in the UK, originally launched in 2000. The programme provides training and education in analysing health data, eHealth, research methods, Information Systems in Health and Public Health Informatics.
- In addition, the Information School delivers an **MSc Data Science** programme, which provides training and education in the analysis of large and complex datasets, including statistical analyses, data and text mining, data visualisation as well as information governance and ethics.
- University of Sheffield leads on Genome Medicine analytics and training through its HEE/Genomics England-sponsored **Genomic Medicine MSc**, a national priority Health Education England commissioned course for NHS Genomic Education. It promotes NHS workforce knowledge, skill and experience to integrate genomics into healthcare.
- TUOS has set up a Wellcome **PhD Doctoral Training** Programme with joint funding available from Wellcome and TUOS for a new cohort of 15 PhD students to work on related topics, in order to make a step change to the field of public health economics and decision science.
- The Insigneo institute is developing a specialised training offer in computational medicine, with its PhD in in silico Medicine and the new **MSc in Computational Medicine**, designed to train specialists in subject-specific modelling to be recruited by research hospitals, regulatory agencies, software developers and biomedical industries.

Public Involvement/Engagement

- Patient and Public Involvement (PPI) is embedded into the research ethos of the [Faculty of Medicine and ScHARR](#). They host a number of PPI groups who participate in providing peer review, feedback, oversight and participation in the design and conduct of our research studies. Most notably these include the [Sheffield Emergency Care Forum](#), the first of its kind in the UK, formed over 10 years ago and the [Sheffield Addiction Recovery Research Panel](#) formed more recently to address the challenges in involving some 'hard to reach' groups in research. We have a track record of innovative PPI in support of mental health studies, based on a longstanding partnership with the Mental Health Foundation, a leading UK mental

health charity. Local PPI partnerships exist with Sunrise, a service user-led group hosted by Sheffield Health and Social Care Trust, and with Sheffield Flourish, a social enterprise supporting numerous user-led recovery projects. For projects involving pulmonary hypertension, we work closely with the Patient's Association (PHA-UK) to develop research projects relevant to this illness. Respiratory representation is also being increased in the Sheffield Teaching Hospitals PPI panel through specific targeted recruitment

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- The Farr@HeRC (led and based in the University of Manchester) – award partners include the Universities of: Bradford, Lancaster, Newcastle, Sheffield and York.
- N8 Research Partnership – Universities of: Durham, Lancaster, Leeds, Liverpool, Manchester, Newcastle, Sheffield, and York)
- Northern Health Science Alliance
- BBSRC Multiscale Biology Network, which coordinates all UK research in multiscale biology across domains of application NHS and other government organisations

DH and NHS

- NIHR Sheffield BRC
- Sheffield Teaching Hospitals NHS Foundation Trust
- NIHR Yorkshire & Humber CLAHRC

Data providers

- NHS Trusts nationally
- NHS Digital
- Public Health England

Industry

- Tech: Avicenna Alliance

International

- Virtual Physiological Human Institute (*Prof Viceconti is current President of this organisation*), which is an international not-for-profit organisation that represents the computational physiology community worldwide.
- Big Data Value Association an international not-for-profit organisation that is implementing the Public Private Partnership between the European Commission and all public and private
- GEOMED, an international, interdisciplinary conference series on spatial statistics, geographical epidemiology and geographical aspects of public health first established in 1997;
- The US Cure Alzheimer's foundation collaboration TUOS with Harvard and Massachusetts Institute for Technology (MIT) for the development of a cross-cutting platform for understanding of the processes of drug treatment and genomic prediction of prognosis for progression.

GEOGRAPHIC REGION: ENGLAND – SOUTH

- Statements of Interest (Sols) to become a Substantive HDR UK Site were received from the following five ROs within this geographic cluster:
 1. **University of Oxford**
 2. **University of Bristol**
 3. **University of Southampton**
 4. **University of Surrey**
- Due to time/resource constraints, this report contains information only on ROs which expressed interests in becoming HDR UK Substantive Sites. The **University of Portsmouth** and the **University of Sussex**, which are based in this geographic region, had also submitted Sols to become a Partner HDR UK but this report doesn't contain any information on their informatics/health data research activities.

Cross-region Strategic Activities

- The NIHR Bristol Biomedical Research Centre (BRC) is in the process of joining the NIHR Health Informatics Collaborative (HIC) programme (see *section 2.2.2*), which is led by the University of Oxford and the Oxford BRC.

University of Oxford

Top-level insights

- The University of Oxford has an international reputation for generating, curating, analysing and sharing very large clinical and genomic datasets and population cohorts, working in partnership with academic, health and commercial organizations across the UK and internationally. Examples include the Wellcome Trust Case-Control Consortium, the Million Women Study, and a series of landmark clinical trials. Over recent years, Oxford researchers have been at the heart of major national projects, including UK Biobank, the 100,000 Genomes Project and the NIHR HIC.
- The University of Oxford Sol focuses on the Big Data Institute (BDI) (housed in a £47M building opened in May 2017, including contribution from an MRC Medical Bioinformatics award), which lies at the heart of Oxford's ambitions in health data science. In addition to this MRC award, BDI has received funding from other diverse sources including the UK Government: UK Research Infrastructure Investment Partnership (£10M); Charity: BHF (£1M); and Philanthropy: Robertson Foundation (£5M capital; £5M fellowships); and Li Ka Shing (£20M fellowships).
- BDI is a multi-disciplinary endeavour that brings together researchers from across the breadth of biomedical sciences and co-locates them with quantitative scientists from the Departments of Statistics, Computer Science and Engineering and the Wellcome Centre for Ethics and Humanities. Further staff recruitment has been underway over the past year.
- BDI is already acting as a central catalyst for large-scale medical research. There is an opportunity to closely align with HDR UK's vision and strategy especially in the areas of clinical informatics for biomedical research, clinical trials and cohort studies, health data analytics, and training.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

University of Oxford's major research investments in biomedical informatics include:

- the **MRC Medical Bioinformatics** award provided £6M capital funding towards the building of the **Big Data Institute** (BDI; *Director: Professor Gil McVean and Deputy Director: Professor Martin Landray*). In addition, the University has invested substantial funds in this endeavour, matching £23M of government, charity and philanthropic funding with £24M from its own capital fund. Generous donations from the Li Ka Shing and Robertson Foundations have provided £25M for senior academic staff and post-doctoral fellowships.
- The **NIHR Oxford BRC** is not only one of the *NIHR HIC* (see section 2.2.2) sites, but it is also where the HIC coordinating team is based (see also section on 'NHS links' below).
- The University of Oxford is a founding partner of the **Alan Turing Institute** (ATI). Areas of data science research at Oxford at the ATI range from machine learning and computability theory to ethics and the social implications of new computational technologies.

Knowledge/Expertise

- The BDI comprise research groups drawn from a wide range of UoO departments (e.g. statistics, computer science, engineering and social sciences). This provides a strong skills-base in the governance, acquisition, processing and analysis of large, complex, heterogeneous biomedical and health datasets for research but also in related ethical and legal issues.
- BDI researchers have experience of collaborations on national flagship programmes including:
 - **UK Biobank** – provided the UK Biobank health informatics hub and data repository since its inception; developing systems for recruitments, clinic assessments, multi-modal imaging and linkage to routine data)
 - **NIHR HIC** – developed a secure, data sharing framework, spanning multiple therapeutic areas across the current network of five BRCs. The HIC approach is metadata-driven (catalogue available via the [NIHR Health Data Finder](#), which Oxford also developed, enabling researchers to more easily compare data across HIC centres.
 - **100,000 Genomes project** – the BRC Clinical Informatics team developed, designed, and oversaw the implementation of a comprehensive clinical information infrastructure for this project.
 - the **Alan Turing Institute** (Holmes is a faculty member), with interests in information security, machine learning, and high-dimensional analysis.

Data

- BDI has access to large rich datasets through its connections with the University of Oxford and other national and international collaborations. For example:
 - Hosting the health informatics hub and primary data repository for **UK Biobank** (0.5M people; 5 years' follow-up; extensive genetic, environmental, functional and health outcomes data).
 - **NIHR HIC viral hepatitis dataset**: anonymised data from patients (currently 1,200 patients) with a diagnosis of hepatitis B, C, D, E or autoimmune hepatitis, collected from the collaborating sites since 2013, primarily from hospital outpatient clinics. These are secondary care – local deep episodic (per consultant episode) data including admission, treatment, follow up, treatment outcome data.
 - **China Kadoorie Biobank** (0.5M people; 10 years' follow-up; linked to death and hospital registries; 300 candidate genotypes; genotyping on 100K ongoing).
 - **Million Women Study** (1.3M women; 15 years' follow-up; linked to death, cancer, hospital registries and CPRD; genotyping in 50K).
 - **Mexico City Prospective Cohort Study** (150K people; 15 years' follow-up; linked to death records; blood and genetic material stored; survey and linkage to hospital records planned).
 - A series of **large clinical trial datasets**, with clinical data (e.g. 100,000 with cardiovascular disease; 10,000 with renal disease; 20,000 maternal health).
 - **Local and national hospitalization, birth and death records for Oxfordshire** (4M people since 1963) and all-England (50M people unlinked 1968-1985, linked since 1998, 7M mother-child pairs).

Translational and innovation capabilities

Industry links

- BDI have collaborative research programmes with several major pharmaceutical companies (e.g. GSK, Merck, Novartis, Novo Nordisk, and Pfizer) and the technology sector (e.g. Apple, Google, Microsoft, Proteus Digital health and Philips).

NHS links

NIHR Biomedical Research Centre

- The **NIHR Oxford BRC** (£113.7M; *Director: Professor Keith Channon*) is a partnership between the University and the Oxford University Hospitals NHS Foundation Trust. Its overarching aims are to enable large-scale data analysis in genomics, imaging and infection, leveraging the statistical, computational and analytical platforms of the BDI and to link with NHS clinical data via the NIHR HIC.

The BRC aims to establish clinical informatics systems, information governance framework, computational infrastructure, and data discovery tools to enable analytical researchers to develop and apply algorithms to selected BRC datasets (e.g. cancer, gastroenterology & mucosal immunity). In the longer-term, there are plans to roll-out clinical informatics and information governance tools across the AHSN and other BRCs; develop a framework for benchmarking and validating digital phenotyping algorithms; extend analytic approaches with other themes.

The BRC informatics team is housed in the BDI.

Academic Health Science Centre

- University of Oxford is also a partner of the **Oxford AHSC** together with Oxford Brookes University, Oxford Health NHS Foundation Trust and Oxford University Hospitals NHS Trusts. Informatics activities within the AHSC are aligned with BDI, and *Professor Martin Landray* (BDI's Deputy Director) chairs the Oxford AHSC Big Data and Clinical Informatics Committee, which provides scientific coordination of activities across all partners, including the BRC. Examples include the development of an integrated, longitudinal research record and data warehousing infrastructure for cancer and microbiology translational research; the delivery of the Oxford NHS Genomic Medicine Centre and of software infrastructure to support large scale clinical genomic research including linkage to national datasets for the 100k Genome Project.

NHS 100,000 genomes project | Genomics England Ltd

- The Oxford team provided informatics support for the **Oxford NHS Molecular Diagnostics Centre (MDC)** and related activities, including a HICF-funded project on the translation of whole genome sequencing (WGS) into clinical practice and the NHS-funded Oxford Genomic Medicine Centre (GMC). These have informed the design and facilitated the delivery of the 100,000 Genomes Project. The team developed data capture and integrated reporting solutions to support the deployment of a 46-gene panel for cancer as an NHS service, providing valuable, timely results on 1406 patients to date. The Oxford GMC was the first centre to offer patients WGS across the full range of eligible cancers, and remains the leading GMC for cancer participation, contributing 35% of the samples to date (*please also see section below on Learning Health Systems*).

NHS Global Digital Exemplar

- The Oxford University Hospitals NHS Foundation Trust recently secured £10M from NHS England as a **Global Digital Exemplar** with plans to collaborate with neighbouring regions and with Cerner sites across the UK. Plans for integration of local primary care (EMIS) and secondary care data (Cerner) are at an advanced stage.

Opportunities

- BDI research themes are in line with HDR UK priorities.
- BDI are enthusiastic about extending the training opportunities through collaboration across the HDR UK. Bringing together experts from quantitative, computational, biological and clinical backgrounds, the BDI is well placed to train the founder generation of biomedical data scientists. Furthermore, the training will benefit from the excellence of established programmes in medicine, statistics, computer science and engineering. For example:
 - BDI can provide space, access to data and compute resource, mentorship and collaborative opportunities for post-doctoral fellows. They are particularly keen to host and develop fellows with unconventional academic careers (e.g. extensive time spent in industry).
 - They also plan to establish a 4 year PhD programme in Biomedical Data Science, largely recruiting students with mathematical and computational backgrounds. This builds on the MRC/EPSRC Oxford-Warwick Statistics Programme (OxW aSP), a 1 plus3 year scheme in the theory, methods and applications of Statistical Science for data-intensive environments and large-scale models.
 - Opportunities for placements in industry or the health sector will be provided.

Potential gaps

- Although BDI has strong analytical expertise in genomics, cohort and clinical trial data, they do not seem to have significant activity on methodological research for the development of learning health systems.
- BDI's e-infrastructure seems to be at the planning stage and there does not seem to be dedicated Data Centre in place yet.

Other relevant capabilities

Research themes

Discovery Science

- BDI (*McVean, Holmes, Kong, Lindgren, Wedge*) capitalises on advances in genomics, immune profile, digital pathology and other laboratory measurements through its connections with the University of Oxford's: Wellcome Trust Centre for Human Genetics; Ludwig Institute and Kennedy Institute; and beyond (Broad Institute, Sanger Institute, EMBL-EBI) to investigate the molecular basis of complex traits, cancer and rare diseases that will lead to new disease pathways and potential treatment targets.
- In collaboration with the Institute of Biomedical Engineering, Oxford Centre for Magnetic Resonance and FMRIB, the BDI (*Noble, Smith*) develops automated quantitative methods for extracting imaging phenotypes from raw medical imaging data, including MRI of the brain in the 100,000 person UK Biobank imaging study.

Precision medicine

Oxford has an interest in this area and building on their current experience in hepatitis C, colorectal cancer, inflammatory bowel disease, asthma. For example, current activity includes:

- **STOP-HCV – Stratified Medicine to Optimise Treatment for Hepatitis C Virus Infection** (MRC Stratified Medicine Consortium £4.1M; *Professor Ellie Barnes*): integrating host and viral genetics to determine why, in an era of new and highly effective (but expensive) drug therapies, some Hepatitis C infected individuals still fail to respond to treatment

- **S-CORT – Stratification in COloRectal cancer** (MRC c. £2.5M; CRUK c. £2.5M, *Professor Tim Maughan*): aims to develop a clinical test(s) that could be given to all colorectal cancer patients upon diagnosis to allow them to have a tailor-made treatment plan, which is guided by their individual biological makeup.
- *Professor Chris Holmes* provides key expertise in statistical analysis of data generated by the stratified medicine consortia.

Population and Public health

- Working in collaboration with UK Biobank (including Edinburgh and Cambridge Universities and the Farr Institute) and with Oxford's large clinical trials, BDI (*Allen, Davies, Lay*) have developed an extensible approach that enables secure transfer and integrated management of data from routine national sources (e.g. HES, cancer, primary care systems) for research cohorts.
- MRC Hubs for Trials Methodology Research, BDI researchers have developed efficient systems for recruitment, data collection, management, quality assurance (including statistical monitoring), and analysis of very large clinical trials, including SHARP and REVEAL. These systems have been used for trials totalling over 65,000 participants and have been inspected by regulatory authorities including FDA, EMA, MHRA.

Learning Health Systems

- The **UK Biobank** and the **Million Women Study** are examples of cohort studies linked to routinely collected EHR data (diagnoses, interventions, mortality, disease registries, primary care).
- Oxford has the **longest population-based record linkage dataset** in the country, linking birth, hospitalization, and death records for Oxfordshire (4M people since 1963) and all-England (50M people unlinked 1968-1985, linked since 1998, 7M mother-child pairs) enabling disease-association studies and analyses of long-term trends in clinical outcomes (e.g. a recent study of dialysis-associated mortality over the past 5 decades)
- Other work that combines routine EHR data with other data sources to improve health includes:
 - the **Modernising Medical Microbiology programme** (funded by NIHR and Wellcome Trust) is supervising the redesign of the processing pipeline for whole genome sequence (WGS) data; establishing data standards and tools for the acquisition of clinical and laboratory information from across a network of centres; providing enabling technology for reporting and surveillance activities. It is envisaged that this will result in an estimated 30% cost reduction with significant greater savings to be realised with further advances in sequencing technology. A national implementation is being piloted for the Public Health England (PHE) mycobacteriology diagnostic service; WGS has been adopted for surveillance and outbreak management for *C. difficile*; and the Oxford BRC team are working with the Stop-HCV consortium to deliver a stratified approach to the treatment of viral hepatitis using new antiviral therapies. This work has been informed by that of the NIHR Health Informatics Collaborative, where the Oxford team have gathered data on cases from across the five comprehensive BRCs, and the development of a national surveillance system for Hepatitis-C with PHE.
 - **Molecular Diagnostics:** Embedding genomic medicine within routine care offers significant patient benefit, particularly for patients with cancer or rare, inherited disorders. The Oxford BRC team (see *also links with the NHS section above*) provided informatics support to the Oxford NHS Molecular Diagnostics Centre and the NHS-funded Oxford Genomic Medicine Centre (GMC). For example, the team developed data capture and integrated reporting solutions to support the deployment of a 46-gene panel for cancer as an NHS service, providing valuable, timely results on 1406 patients to date. The Oxford GMC was the first centre to offer patients WGS across the full range of eligible cancers, and remains the leading GMC for cancer participation, contributing 35% of the samples to

date. The team have designed and developed support for multi-disciplinary team (MDT) working in clinical genetics: the **Oxford Genomic Medicine MDT** has been considered a model for embedding genomic medicine in NHS clinical practice. Since its implementation, referrals have increased 300%, with 30% of the new referrals coming from departments outside clinical genetics. >500 individuals have been offered and have accepted WGS. Based upon the pilot programme experience, 25% of these may expect a new, validated diagnosis. The team have also developed the data specifications and the software for data acquisition and management used in the 100,000 Genomes Project.

- The **Oxford BRC Clinical Informatics team** provides supports work on a diversity of research themes and disease areas that seek to: recruit patients for particular studies; assess changes in health and disease over time; evaluate the impact of new diagnostics or interventions; or develop new approaches to disease management. The BRC Clinical Informatics team provides a scalable, configurable, multi-directional information flow across organizations (including primary and secondary care) and between patients, healthcare professionals and researchers. This work builds on existing strong collaborations with our **Digital Health and Imaging programmes**. Examples include:
 - ◆ digital health approaches to management of chronic disease using EHR, smartphone-enabled sensors/cameras, EHR and machine learning
 - ◆ development of clinically applicable algorithms for patient stratification in inflammatory bowel disease
 - ◆ assessing the impact of (often silent) atrial fibrillation on cardiovascular, stroke, and vascular dementia diseases
 - ◆ using combined clinical and genomic data to identify emergence of infection threats
- Randomized controlled trials: Oxford scientists have worked on a number of practice-changing clinical trials including for example, the **3C randomized trial** of immunosuppressive strategies following renal transplantation depending entirely on routinely collected EHR data to determine efficacy and safety of the different interventions over the medium-longterm.

Citizen-driven/Digital health

- The BDI together with the Institute of Biomedical Engineering and Department of Statistics apply machine learning approaches to signal processing and data fusion in order to interpret data from mobile sensors. Examples include: quantifying and characterising activity patterns from 7-day accelerometer data in 100,000 **UK Biobank** participants; and web-cam based assessment of oxygen saturation and vital signs for continuous physiological monitoring.
- BDI (*Hinds, Young*) in collaboration with the Department of Psychiatry (*Gallacher, Lovestone*) are able to assess cognitive and mental health through web- and smartphone-based technologies, including the **Healthy Minds** assessment of cognitive function in 140,000 UK Biobank participants; and the **TrueColours** system for self-monitoring of mood disorders in bipolar disorder.

Disease areas

- University of Oxford has research programmes on a number of disease areas with particular focus on complex disorders (e.g. obesity, diabetes, CVD), cancer and infection.

Analytics

BDI statisticians and biomedical engineers (*Fraser, Gething, Holmes, McVean, Rittscher*) are engaged in the development of statistical and machine learning methods to gain new insights from multi-modal, biomedical and healthcare data, including:

- **Clinical phenotyping:** Use of (semi-)automated approaches (including regression models and machine-learning) to derive clinical phenotypes, distinguish disease subtypes, and identify common features across multiple diseases.
- **Longitudinal analysis of patient data:** Development of statistical methods to derive information on response to treatment, progression, relapse or paroxysmal changes in disease state from routine healthcare data.
- **Analysis and integration of high-dimensional, heterogeneous biomedical data:** New approaches that combine biological knowledge with unsupervised machine learning to assess relationships between diverse data sources (e.g. combined genomic, imaging, and clinical data).
- **Epidemiology:** Reliable analytic approaches to genetic epidemiology, prospective cohort studies, large randomized trials and meta-analyses.
- **Geo-spatial and temporal analyses:** Mapping and modelling infectious disease and drug-resistant infection.

The Oxford BRC Big Data analytics sub-theme allows researchers from across the BRC to apply HPC to patient-derived research data efficiently and at large scale. Using exemplars from other BRC themes, the team is developing analytical methods for large, heterogeneous datasets that can be applied to similar scientific problems both locally and nationally. Methods (including Bayesian and machine learning approaches) will be developed to support analysis of high-dimensional biomarkers (immune profiling, functional genomics, imaging) in areas such as *Cancer, Gastroenterology and Mucosal Immunity, Molecular Diagnostics, Genomics, Imaging*. Methods to characterise clinical phenotype – including disease subtypes or commonality between conventionally “different” diseases – and the influence of treatment or other external triggers on their severity over time – will benefit proposed translational medicine projects within *Diabetes, Cardiovascular, Respiratory, Stroke and Vascular Dementia, Multi-morbidity and Long-term Conditions, Musculoskeletal, Gastroenterology and Mucosal Immunity, Cancer, Antimicrobial Resistance and Modernising Microbiology*.

e-Infrastructure

Compute

- The BDI currently *has provision* for 32 racks with cooling and 320kW of power, sufficient for around 60PB of storage and over 8,000 cores. The relative allocation to storage and compute will be determined through more detailed planning. The initial facility (~£1.5M of hardware) will be expanded to reach capacity over the coming 18-36 months.
- Over the next 1-2 years, the BDI will define and implement a strategy for development of off-site HPC capability.

Training

- Oxford has a number of taught **Masters programmes**, including short courses in software engineering which can be combined to obtain an MSc in Software & Systems Security (accredited by GCHQ).

- In collaboration with the University of Warwick, it has **MRC/EPSRC Centre for Doctoral Training in Next Generation Statistical Science Programme (OxWaSP)**. The Programme aims to train graduate statisticians in the theory, methods and applications of statistical science for data-intensive environments and large-scale models.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@CIPHER (led and based in Swansea University) – award partners include the Universities of: Cardiff, Bristol, Brighton, Oxford, Exeter and Surrey, Western Australia, Monash, Ottawa; and the Welsh Government; and the Finnish National Public Health.
- MRC Medical Bioinformatics capital award in the BDI
- UK Biobank
- DPUK – DRI
- Alan Turing Institute

DH and NHS

- NIHR HIC
- NIHR Oxford BRC
- Oxford AHSC
- Oxford AHSN

Data providers

- Primary care systems suppliers: EMIS, TPP, INPS
- NHS Digital
- Public Health England

Industry

- Pharma: GSK, Merck, Novartis, Novo Nordisk, and Pfizer
- Tech: Apple, Google, Microsoft, Proteus Digital health and Philips

International

- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR
- Global Alliance for Genomics and Health
- NIH eMERGE network (Vanderbilt)
- NIH National Center for Biomedical Ontology and Center for Biomedical Informatics Research
- NIH Human Connectome Project
- US Precision Medicine Initiative
- US Institute of Health Metrics and Evaluation
- other collaborative medical research projects and programmes across the developed and developing world

University of Bristol

Top-level insights

- University of Bristol Sol focuses on its existing data assets that include one of the largest and most deeply genotyped and phenotyped four-generation population cohort in the world (Avon Longitudinal Study of Parents and Children, ALSPAC) with linkage through to social and health care data. Developing infrastructure to curate and share these data and address problems of cost efficient follow-up and participation bias in exemplar projects are the focus of their contribution to Farr@CIPHER.
- A key strength of the University of Bristol is in Digital Health with expertise not only in the acquisition and interpretation of health data from personal devices but also pioneering the use of Internet of Things for health research with several studies deploying wirelessly networked devices in local homes. Over the next 3-5 years the University of Bristol plans to create a physical hub for digital health research, part of a £300M investment in a second campus. Bristol is Europe's most advanced smart city with sensors across the city to track energy usage, air quality and traffic flows, informing an innovative operating system to monitor and deliver the city's needs. The University Hospital Bristol Foundation Trust has earned a Global Digital Exemplar status.
- The newly-funded NIHR Bristol BRC has a theme on 'Biostatistics, Evidence Synthesis and Informatics' aiming to use cutting-edge approaches to organise and analyse different information types to inform clinical decisions.
- The University of Bristol Advanced Computing Research Centre is one of the country's leading centres for HPC. BlueCrystal, the University's flagship HPC supercomputer, is the second most powerful of any UK University and ranked 301st in the world in 2016.
- As part of the GW4 Alliance, the University of Bristol works closely with its partners: Cardiff, Bath and Exeter Universities. Other collaborations include a recently awarded BRC with a foundation in population health and a focus on scientific discovery through analyses of large and complex datasets.
- Their assets represent over £100M external investment matched to a high degree by institutional support.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- High-level institutional commitment in data science research, which is demonstrated by investments such as:
 - the **Jean Golding Institute for Data-Intensive Research**, which brings together research in data analytics and analysis methods, data infrastructure and applications across all UoB Faculties;
 - the joint Wellcome/ University of Bristol -funded **Elizabeth Blackwell Institute for Health Research**, which fosters collaborations between world-leading researchers who work on data-intensive research in statistical bioinformatics; clinical, translational, and infrastructural epidemiology; and computational genomics;
 - the **Brigstow Institute**, which brings researchers from different disciplines together with a range of partners to experiment in new ways of living and being. Living well with technologies is a key theme.

- Other informatics related investments include the MRC **Integrative Epidemiology Unit** (£23M funded by MRC and University of Bristol), which is linking ‘omics, population health and environmental data to investigate the underlying causes of disease, and with a strong cross-cutting theme in bioinformatics and data mining developing novel methodology and tools.

Knowledge/Expertise

- Since 2016, the University of Bristol has appointed four new Chairs in Data Science (*Bullock, Salas, Dowsey and Wood*) alongside new lecturers in Human Computer Interaction for Health and in Computational Health Behaviour Analysis.
- Expertise in acquisition and interpretation of health data from personal devices is a key strength. The University of Bristol has developed key competences in machine learning from these sources of time-series data, tools to conduct large scale digital health trials, related research ethics processes, participatory design and public engagement.

Data

Some of the many data assets within the University of Bristol include:

- The world’s most data-rich pregnancy cohort (currently supported by an MRC/Wellcome strategic award (2014-19)). The £1.8M Wellcome-funded **Project to Enhance ALSPAC through Record Linkage** (PEARL) gained permissions to link to routine health and social data on >90% of original ALSPAC participants (98% for health data). Around 80% of participants have linkage to primary care records.
- The University of Bristol co-leads the **Born in Bradford** (BiB) pregnancy cohort (funded by MRC, ESRC, Wellcome and British Heart Foundation), a two-generational deeply phenotyped pregnancy cohort, which has genome-wide genetic and DNA methylation data and NMR metabolomics data on both parents and children. The BiB biobank is located in Bristol.
- The Bristol-based **UK Renal Registry** is linked to local EHRs and laboratory data, which are captured in real-time and made available online to patients across the UK via **PatientView**. This infrastructure is being used to support five data efficient trials of interventions in renal patients.
- The University of Bristol has a high-quality database of genome-wide association study results (containing results of ~4billion genetic association analyses from >1,000 datasets, to be substantially expanded with UK Biobank data).
- The University of Bristol is leading on creation of a **Digital Health Asset Register** of all available data sources held within organisations across the Bristol, North Somerset, and South Gloucestershire region, with support from the Elizabeth Blackwell Institute and Bristol Health Partners. This includes data held by UoB, the University of the West of England, University Hospitals Bristol NHS Trust, North Bristol NHS Trust, Avon and Wiltshire Mental Health Care Trust, the CCGs and Commissioning Support Units and Bristol City Council.

Compute

- The University of Bristol Advanced Computing Research Centre is one of the country’s leading centres for HPC.
- **BlueCrystal**, the University’s flagship HPC supercomputer (£3M) provides access to 16,000 cores, whilst the Research Data Storage Facility provides 2.5PB of storage (soon to be 12PB).

- The **GW4 Alliance**, brings together four of universities in the south west of the UK; the universities of Bath, Bristol, Cardiff and Exeter. Together with Cray Inc. and the Met Office, GW4 Alliance has been awarded £3M by EPSRC to be installed in 2017 and deliver a new Tier 2 HPC service for UK-based scientists. The system will be one of the world's first to be based on Broadcom's Vulcan server-class chip. Details of this device are still under NDA, but the Vulcan CPU is generating excitement because it trades off much greater provision of memory bandwidth for less emphasis on peak FLOP/s, the former being more important for most scientific codes. Providing access to such a machine as a national service should therefore enable the UK's HPC community to quantify the benefit of memory bandwidth focused CPUs, thus informing future system procurements from Tier 1 to Tier 3.

Translational and innovation capabilities

Links with industry

- Partnerships include pharmaceutical companies such as GSK, Biogen, Merck, Sanofi and Ferring. There are also collaborations in place with Roche Diagnostics and Medtronic who access the University's informatics capabilities (e.g.MR-Base) to complement their drug discovery and biomarker platforms and programmes.

Links with NHS

- The newly-funded **NIHR Bristol BRC** (£21M) is a partnership between **University Hospitals Bristol NHS Foundation Trust** and the University of Bristol. This BRC aims to catalyse development of new diagnostics, treatments based around population health. The cross-cutting "Translational Population Science" theme analyses contributions of the genome, epigenome and microbiome to disease risk, to inform prevention and treatment. The also cross-cutting "Biostatistics, Evidence Synthesis and Informatics" theme uses cutting-edge approaches to organise and analyse different information types. A key aim, particularly in the Cardiovascular Disease and Perinatal and Reproductive Health research themes is to harness routine clinical measures to inform clinician/patient discussions around treatment decisions.
- Bristol is in the process of joining the NIHR Health Informatics Collaborative (HIC) programme, and will therefore lead a new project analyzing routine health care data combined across the six participating acute NHS Trusts.
- University Hospitals Bristol NHS Foundation Trust has NHS **Global Digital Exemplar** status (£10M) which provides the opportunity to increasingly link to more diverse datasets and work with industry.
- The University of Bristol works with the **NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) West** and the **West of England AHSN**.

Opportunities

- The University of Bristol has recently approved a Digital Health Strategy, the implementation of which will be led by Ian Craddock and which will involve investments in additional new posts. Its expertise and investment can help catalyse HDR UK research activities in this area.
- The University of Bristol plans to capitalise on the emergence of high-throughput and large-scale 'omics profiling to spearhead molecular phenotyping at scale. A multi-platform molecular phenotyping pipeline, data bank and bioinformatics resource will be established for the **Manchester's Stoller Biomarker Discovery Centre** (£18M) clinical proteomics facility and combined with University of Bristol's established Brainshake NMR metabolomics acquisition and analysis platform that includes complementary mass spectrometry metabolomics, lipidomics and proteomics data from their large-scale cohort studies. The proposed system will realise statistical analyses and mechanistic discovery at unparalleled scale in the 'omics field.

- There are exciting plans to create an **Avon-wide Digital Population Laboratory** including links to **Connecting Care**, the regional initiative to digitally link social and health care data. This will be a data intensive research resource to support the application of large-scale ‘omics and other analyses to the discovery of causes, testing of hypotheses and development of more effective prevention and treatment. It will be distinguished from other such initiatives by the integration of ALSPAC and other cohorts: a relatively stable population with unusually high coverage of linkage to routine data; our expertise in establishing and using data linkage pipelines, sharing linked data and developing and deploying cutting edge sensor technology; and access to a city-wide interactive network provided by **Bristol is Open**, and with the aim of extending coverage of linkage to the full Connecting Care databank.

Potential gaps

- The University of Bristol is not leading on any of the MRC Stratified Medicine Consortia; and it does not appear to be involved in any other stratified /precision medicine initiatives.

Other relevant capabilities

Research themes

Discovery Science

- **Brainshake** is an automated high-throughput serum NMR metabolomics platform that was developed by *Professor Mika Ala-Korpela* and his Computational Medicine team at the University. It provides quantitative molecular information for 233 metabolic measures. Analyses with this platform are commercially available via Nightingale Health Ltd.

Precision medicine

Precision medicine is prominent in the Bristol NIHR BRC led across several work-streams by *Professor George Davey Smith* and his colleagues in the MRC Integrative Epidemiology Unit. Precision medicine initiatives have been further developed by work also involving *Professor Davey Smith* based on the Bristol-based UK Renal Registry – for example:

- the **NURTuRE** project (MRC £3.1M; *Professor Moin Saleem*) will use the valuable samples and clinical data collected in the NURTuRE-NS (NephroS) and NURTuRE-CKD registry, and generate new data and methods that will allow us to redefine the diagnostic categories of nephrotic syndrome and CKD. This will help to provide individualised diagnostic tests and new therapies and will be based on deep analysis of patient DNA, blood and biopsy samples provided within this study. The four-year project is boosted by in-kind industry partner contributions and the continued involvement of Kidney Research UK and the UK Renal Registry.

Other Bristol BRC-led projects in this area include:

- The **BRC Reproductive and Perinatal Health Research Theme** conducts translational research, covering women’s reproductive life from fertility through pregnancy to the menopause, that supports appropriate precision prevention and stratified medicine to improve women’s and their children’s health. In particular, they are determining optimal IVF treatment strategies to increase livebirth rate and maximise short- to long-term health of mothers and infants, and will be testing the feasibility and acceptability of stratified medicine approaches to IVF treatment.
- The **BRC Mental Health Research Theme** is identifying personality traits contributing to treatment response in psychological therapies for depression, establishing biomarker profiles to improve prediction and risk stratification of psychotic experiences, and identifying biomarkers predicting response to psychological therapies for depression.

Population and Public health

- **The PEARL project:** (i) demonstrated that linkage to routine health and social care data can add substantial value to population-based research; (ii) established data pipelines to capture routine health and social data at scale; and (iii) developed associated governance structures and informatics tools.
- Building on the PEARL project by extending linkage to new data, including the 1M local citizens whose records contribute to **Connecting Care**. UoB researchers are exploring governance and data sharing structures to allow secondary research use of these data through the **Reach West** initiative, following the successful Scottish Go-share model. Their ambition is to incorporate 'omic analyses based on residual routine samples into an **Avon-wide Digital Population Laboratory**, and enhancement through the use of sensor technologies.

Citizen-driven/Digital health

- The **Digital Health Engineering** research group particularly focusses on chronic health conditions, for which there is a clear imperative to develop non-intrusive technologies for capturing data on causes, symptoms and exacerbations over long periods of time, in large populations and outside clinical environments. Moving beyond the use of smartphones, they are pioneering use of Internet of Things (IoT) for health research with several studies deploying wirelessly networked devices in local homes. Exemplar projects include:
 - the **SPHERE project** has developed machine learning capability to process time-series data from off the shelf and bespoke devices, wearables, depth cameras and home appliances into health behaviour data. This capability is already the backbone of a number of **Bristol-based digital trials** including an initial “100 Home” deployment of the SPHERE sensor system. Applications will include: characterisation of the recovery of orthopaedic patients at home; and the MRC-funded **CUBOiD project** that is using IoT devices to examine behavioural biomarkers for disease progression related to dementia.
 - the **EPSRC Sensor Platform for HEalthcare in a Residential Environment (SPHERE)** project (£15M) involving clinicians, engineers, designers, and social care professionals in developing remote sensors situated in domestic environments.
 - the **EurValve project** (£4M; H2020) is using sensors to characterise outcomes of heart valve replacement.
 - the development and validation of an online tool that is available and as a smartphone app to predict live-birth success per treatment cycle, and determined the impact of transferring 1, 2 or 3 embryos on live births and perinatal outcomes.

Learning health systems

- Over the last decade Bristol has developed a **learning Intensive Care Unit (ICU) healthcare system**. Working with large datasets collected as part of normal care to use information captured in real time to characterise treatment, inform better clinical decisions, evaluate interventions and predict future changes in the status of patients that are of clinical and operational benefit.

Disease areas

- The University of Bristol has research programmes on a number of disease areas with particular focus on: reproductive health, cancer and CVD.

Analytics

- The **Evidence Hub** is an integrated analytical platform for large-scale systematic epidemiological hypothesis generation and testing.
- The **LD Hub** is a centralized database of summary-level GWAS results for 173 diseases/traits from different publicly available resources/consortia and a web interface that automates the LD score regression analysis pipeline. It was developed collaboratively by Broad Institute of MIT and Harvard and MRC Integrative Epidemiology Unit, University of Bristol.
- The **MR-base** is a database and analytical platform for Mendelian randomization being developed by the MRC Integrative Epidemiology Unit at the University of Bristol.
- The **Melodi literature** mining software exploits a novel graph database containing >26M PubMed records and 18M literature-derived semantic triples to enable mapping of mechanistic pathways. Systematic integration with causal evidence and genetic correlation will enable a step-change in identification of risk factors and potential interventions.
- The **ROBINS-I** tool assesses risk of bias in non-randomized studies of interventions and extending this for use in assessing studies of non-intervention risk factors for disease.
- Through ESRC funding the **Centre for Multilevel modelling** (CMM) has developed an interface to their Stat-JR software: **Statistical Analysis Assistants** (SAA) that produce, from simple user inputs, bespoke analysis documents running through all stages of a statistical analysis from exploratory data analysis through model fitting to checking and residual analysis. This will be adapted to construct **informatics analysis assistants** (IAAs) to make cutting-edge analytical techniques and tools accessible to researchers beyond the community of specialist bioinformaticians and biostatisticians with the required knowledge to work with applied researchers.

e-Infrastructure

Compute

- **BlueCrystal**, the University's flagship HPC supercomputer (£3M) provides 16,000 cores. This facility is used intensively across the University; ~8% of the capacity has been purchased for the MRC Integrative Epidemiology Unit. In addition, the UoB Research Data Storage Facility provides 2.5PB held on RAID 6 devices. The second phase of Research Data Storage Facility (£1.2M) is underway and will provide an additional 12PB of tape storage.
- Through its **GW4 Alliance** membership, the University of Bristol will have access to a new (to be installed in 2017) Tier 2 HPC service for UK-based scientists. The system will be one of the world's first to be based on Broadcom's Vulcan server-class chip.

Training

- The University of Bristol has just launched a new Vice Chancellor's Fellowship scheme in **Digital Innovation and Wellbeing** for early and mid-career researchers.
- In addition to PhD training, there are postgraduate taught courses that include: an **MSc in Advanced Computing – Machine Learning Data Mining and High Performance Computing**;
- Also, short courses in **applied data linkage**; and use of the DataSHIELD infrastructure for non-disclosive analysis of sensitive research data.

Public Involvement/Engagement

- The University of Bristol works closely with the **People in Health West of England (PHWE)** initiative, which promotes innovative and effective public involvement in research and evidence-based service improvement amongst professionals and members of the public. Partners in PHWE include: the University of Bristol, Bristol Health Partners, and NIHR Clinical Research Network-West of England, CLAHRC West and the West of England AHSN.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@CIPHER (led and based in Swansea University) – award partners include the Universities of: Cardiff, Bristol, Brighton, Oxford, Exeter and Surrey, Western Australia, Monash, Ottawa; and the Welsh Government; and the Finnish National Public Health.
- University of Oxford
- GW4 Alliance University partners: Bath, Bristol, Cardiff and Exeter
- University of Manchester – Stoller Biomarker Discovery Centre
- UK Biobank
- ESRC CLOSER consortium
- LSHTM

NHS and other government organisations

- NIHR Bristol BRC and NIHR HIC
- Genomics England
- Bristol Health Partners (Bristol's three NHS Trusts; three CCGs and two Universities)
- West of England AHSN
- NIHR Clinical Research Network-West of England
- CLAHRC West

Data providers

- CPRD
- NHS Digital
- Public Health England

Industry

- Pharma: GSK, Biogen, Merck, Sanofi , Ferring, Roche
- Tech: Medronic

International

- MIT Broad Institute

University of Southampton

Top-level insights

- The University of Southampton strengths build on a number of unique investments in health data science, which are underpinned by a strategic focus on interdisciplinary research that includes: the Interdisciplinary Institute for Life Sciences; the Web Science Institute; the Microsoft Institute for High Performance Computing; the Administrative Data Research Centre for England; and the ESRC Centre for Research Methods.
- These investments are complemented by the University Hospitals Southampton Foundation Trust's (UHS) pioneering activities in health informatics that has won them NHS Global Digital Exemplar status.
- The University of Southampton activities in informatics research are becoming even more clinically relevant with the establishment of an NIHR Biomedical Research Centre with a cross-cutting theme in data analytics and an ambition to invest in shared space at the hospital for secure access by researchers to NHS data. Furthermore, the University of Southampton works closely with the Wessex CLAHRC and AHSN and have identified digital tools and services as key areas of collaboration.
- The University of Southampton has one of only nine specialist Microsoft Research Institutes helping define the information security capabilities. This is underpinned by its status as an academic centre for excellence in cyber security research awarded by GCHQ.
- The University of Southampton has not been part of the Farr Institute. However, there are strong links and collaborations with the Farr@Scotland through Professor Jeremy Wyatt, who is Professor of Digital Healthcare, and the Director of the Wessex Institute in Southampton since 2016, and he used to be the Director of the Dundee Health Informatics Centre (2005-2010). But also to the Leeds Medical Bioinformatics award as he was Chair in eHealth Research at the University of Leeds, and one of the lead investigators on the health informatics and clinical epidemiology workstream of this award prior to his Southampton appointment.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

The University of Southampton's major interdisciplinary investments in the areas of biomedical research and population health with relevance to health data science include:

- The **Interdisciplinary Institute for Life Sciences** (£65M; *Director: Peter J. Smith*) brings together an interdisciplinary collaborative network of more than 300 members representing all University Faculties. One of four research priorities in biomedicine is **The Human Nexus**, a uniquely 21st century global challenge drawing together data from biomedical and environmental sources with cutting edge developments around the semantic web, Web Observatory, modelling and high-end computing, to evolve new, better ways of understanding complex data as they impact on disease definitions and the development of personalised and precision treatments.
- The **Wessex Institute** (240 staff, *Director: Professor Jeremy Wyatt*) provides high-quality, practice-based research to support decision makers in health and healthcare.

- The **Microsoft Institute for High Performance Computing** (*Director: Professor Simon Cox*) was created in November 2005. It is the only one of its kind in the UK (the only other European Institute is in Stuttgart) and one of a total of nine institutes worldwide. The Institute pushes state-of-the-art technologies to tackle real-world scientific and engineering problems and works closely with industrial partners that span the aerospace, automotive, bioengineering, marine and telecommunications sectors.
- The **Web Science Institute** (£9M, founders *Dame Wendy Hall & Sir Nigel Shadbolt*) brings together world-leading interdisciplinary expertise to tackle the most pressing global challenges facing the World Wide Web and wider society today. It is necessarily interdisciplinary, as much about social and organisational behaviour, as about the underpinning technology. Members include researchers from the social and computational sciences, the humanities, medicine and health sciences, business and law and the natural sciences.
- The **Administrative Data Research Centre for England** (ADRC-E) (£12M; *Director: Professor Peter Smith*) (see section 2.5.1) is hosted at the University of Southampton with UCL as a key partner. ESRC has invested in safe settings and secure data lab technical infrastructure in both sites.
- The ESRC-funded **National Research Methods Centre** comprises a collaborative partnership, between the Universities of Southampton, Manchester, and Edinburgh. It runs a programme of methodological research for the analysis of complex data with linked and time-dependent structures, enabling social scientists to address key substantive and policy-relevant research questions.
- The University of Southampton has provided expertise, practical advice and in the very early stages, seed funding to the **Open Data Institute** (ODI) (£10M). It has an honorary founding partner status and the ODI's Chairman and co-founder, Sir Nigel Shadbolt, was Professor of Artificial Intelligence at the University of Southampton before moving to Oxford in 2017.

Knowledge/Expertise

- The University of Southampton has strengths in computing and mathematical sciences, including data analytics, web sciences, and cyber security; as well as engineering and statistics applied to social sciences.
- In addition, it has particular expertise in precision medicine, life course epidemiology and translational public health and health technology assessment.

Data

- Research data at the University of Southampton includes longitudinal, observational and interventional research cohorts. Examples include: the **Southampton Women's Survey** (n=12500), the **Hertfordshire Cohort Study** (n=3000), **Isle of Wight Birth cohort** (n=1400) and **Isle of Wight 3rd Generation cohort** (n=400), the **maternal vitamin D osteoporosis study** (n=1431), the **Southampton assisted conception study** (n=750), **UK pregnancies better eating and activity trial** (n=3000) and the **Motherwell cohort study** (n=950). As well as deep longitudinal phenotyping, the cohorts include high-dimensional data such as imaging and multi-level 'omics including genetics, epigenomics, transcriptomics, proteomics, metabolomics and metagenomics.
- The University Hospitals Southampton (UHS) NHS Foundation Trust has invested in a new **NHS EPR**, containing structured patient-centric data. A constantly updated **non-operational copy of UHS data** is available for research purposes. Primary and social care data within Hampshire is available through the **Hampshire Health Record** (HHR). Currently 80% of GP surgeries within Hampshire clinical commissioning groups deposit patient data in the HHR, as do hospitals and community care teams.
- A collaboration of the University of Southampton with the **Ordnance Survey** and local authority partners provides access to local environmental, socioeconomic and spatial data to better understand disease and target services.

Compute

- The University of Southampton houses **IRIDIS 5**, one of the most powerful academic supercomputers in England on installation, a £5M investment with 18,000 cores and a peak of over 1Petaflop/s along with 2.5Pb of online storage and 10Pb of deep offline store.
- The University of Southampton are early users for the **Azure Cloud Compute** resource and fostered a unique relationship with Microsoft Research to become one of only 12 specialist institutes helping define the information security capabilities of Azure, as well as receiving GCHQ recognition for cyber security research.

Translational and innovation capabilities

Links with industry

- The University of Southampton has fostered a unique relationship with **Microsoft Research** and has been identified as one of only 12 specialist institutes (including government agencies) helping define the information security capabilities. This is underpinned by its status as an academic centre for excellence in cyber security research awarded by GCHQ.
- Other collaborations include: my mHealth (*Wilkinson*); BC Platforms; IBM Hursley; The Learning Clinic; and Volv Goba.

Links with NHS

- The **NIHR Southampton BRC** (£14.5M) is a partnership between the University Hospitals Southampton NHS Foundation Trust and the University of Southampton and has a cross-cutting theme in data analytics.
- The University of Southampton **Wessex Investigational Sciences Hub laboratory** (WISH Lab) provides a bespoke laboratory for translational research requiring quality assurance to regulatory standards. It also part of the Genomics England initiative and hosts the **Wessex NHS Genomic Medicine Centre** (£2M), which is working on secure, audited pipelines to process genomics data in a manner suitable for clinical decision making.
- The University Hospitals Southampton(UHS) Foundation Trust was successful in their bid to become a **Global Centre of Digital Excellence** (£10M), chosen to deliver innovative and pioneering approaches to their digital services. UHS is recognised as a leader in NHS IT having overseen the development of a number of pioneering projects in recent years. These include the development of **My Medical Record** (a personal online patient health record), electronic blood tracking via wristbands, a mobile vital signs early warning system and a full electronic record in all critical care units. In addition, the trust has more than 10M documents in its electronic patient records system and runs a novel **eDischarge summary programme** which enables clinical documents to be sent electronically.
- The **NIHR Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Wessex Data Science Team** (*Director: Professor Anne Rogers*) have been looking at data modelling and simulation to try and improve the efficiency of community health care.
- **The Wessex AHSN** is working in partnership with the University of Southampton and UHS Trust to deliver its programme **on Digital Health and Optimising Intelligence** (population targeting).

Opportunities

- The University of Southampton is investing further in their informatics research, and making it even more clinically relevant, with a planned joint University-NHS £1.2M investment in a **Biomedical Informatics Centre** on the hospital site, in which computer, clinical and data scientists can safely and securely access clinical, imaging, laboratory and research datasets through the NHS N3 network. This will deliver a powerful boost to Southampton's inter-disciplinary research capabilities.

Potential gaps

- The University of Southampton has unique niche expertise in in basic and translational health data research.

Other relevant capabilities

Research themes

Discovery Science

- The University of Southampton supports research (*Lillycrop, Bell, Holloway*) on: epigenomic and metagenomic mechanisms for the transmission of environmental influences to risk of non-communicable disease; and molecular correlates of disease serving as prognostic and stratification markers, and surrogate endpoints for drug discovery.
- University of Southampton researchers (*Djukanovic*) are participating in the **EU IMI project** (€22M) entitled **U-BIOPRED** (Unbiased BIOMarkers for Prediction of Respiratory Disease), linking 'omics to clinical data to better understand severe asthma.

Precision medicine

- The **Southampton CR UK Centre** (*Johnson*) is a key enabler of using 'omics analysis for stratified approaches to oncology.
- The Southampton **Genomic Informatics Group** (lead: *Ennis*) works closely with colleagues in Genetic Epidemiology (*Collins*) in the analysis of next generation sequencing data and its clinical applications.
- Other research activities in stratified medicine include clinical genotyping, consent and sample taking methods and medical data input to the genetic and epigenetic basis of childhood developmental disorders and asthma (*Temple, Djukanovic*).

Population and Public health

- The **MRC Life-course Epidemiology Unit** (£15M, *Cooper*) undertakes research that addresses environmental influences throughout the life-course on chronic musculoskeletal and metabolic disorders and aims to develop population-based and high-risk preventive strategies against these. The Unit maintains and develops long-term cohort studies, which it uses to develop and test interventions designed to improve health and health behaviours. The Unit has also a **Centre for Participatory Medicine** (*Barker*).
- University of Southampton researchers (*Roderick, Sheron*) conduct research into community-based identification from routine data of people with liver and kidney disease using linked datasets for risk profiling and evaluating health care, in particular primary care records integrated through the Hampshire Health Record.
- University of Southampton researchers (*Wyatt*) study data coding, for example of: GP allergy data in 2M encounters (with Edinburgh); and family cancer (with Leeds), using a new method to measure and improve the accuracy of HES diagnostic codes (with Leeds & U Paris V).

Citizen-driven/Digital health

- University of Southampton research in this area includes: study GP responses to citizen-collected data (*Giordano*); real-time collection and visualisation of glucose levels in paediatric type 1 diabetics (*Guy*); and validation of SMS messages to capture breast feeding data in 350 women in Tayside (*Wyatt* – in collaboration with Dundee); citizen-use of personal health records (eg. systematic review of online access to personal health records with RCGP) and the accuracy and classification of, and privacy and other risks posed by, health apps for citizens (with Oxford).
- NHS / industry R&D includes a suite of CE-marked self-management apps developed by My mHealth based on UHSFT research (*Wilkinson*), now used in 15+ CCGs by tens of thousands of patients.
- *Professor Jeremy Wyatt* co-organised with *Professor Susan Michie* (UCL) the 2015 MRC / NIH consensus workshop on development & evaluation of digital interventions. He also works on app and algorithm quality with RCP, BSI, NICE, CQC & MHRA.

Learning health systems

- University of Southampton researchers (*Wyatt*) study clinical data use, novel data formats and clinical decisions (e.g. RCT in 200 women with breast cancer, with Regensburg), novel decision support methods (e.g. cluster RCTs of decision support to help GPs use ACOVE indicators in frail elderly with Amsterdam; systematic review and empirical study of automation bias with UBC Victoria).
- They also explore the challenges of causal inference from routine data (e.g. study of MI in 9,600 CPRD records, showing failure of propensity scoring with Imperial College and Dundee) and obtained CSO funding to validate regression discontinuity methods in 12,000 older women with breast cancer (with Imperial College and Edinburgh).

Disease areas

- The University of Southampton has expertise in disease areas including: cancer immunology; respiratory, allergy and critical care; life-course diseases, musculoskeletal disease, nutrition and metabolism; and infection.

Analytics

- The **Human Nexus** is a collaborative project that brings together interdisciplinary expertise from the University of Southampton Institute for Life Sciences, the Web Science Institute, and Southampton Centre for Biomedical Research. The aim to integrate knowledge elements of research into one system, bridging small systems (e.g. cell signalling), intermediate systems (e.g. healthy and diseased humans) and large systems (e.g. society and environment). The tools and analytics developed in this project will have broad application, but not just to disease states, such as cancer and diabetes, but also to environmental factors and climate change.
- The **Clinical Informatics Research Group** (*Batchelor*) developed the **EDGE Software**, which are cloud-based clinical research tools used by 70% of UK CTUs, giving real-time views of research activity in a single hospital or research network. Manages over 100,000 research projects and holds 4.2 million patient records. over 30,000 listed users.
- Other methodological research at the University of Southampton includes: have development of novel machine learning algorithms to support severe asthma biomarker discovery using proteomic and lipidomic methods (*Brodzki*); development and testing of innovative software for clinical image processing in respiratory and neurological diseases (*Bennet and Guy*); using information design techniques to display breast cancer patient data on a timeline to improve clinical decisions (*Rew*).

e-Infrastructure

Compute

- The University of Southampton houses IRIDIS 5, one of the most powerful academic supercomputers in England on installation, with >8,000 cores installed in 2017/18.

Trusted Research Environments

- The University of Southampton has established an NHS information governance toolkit compliant data lake for storage and analysis of patient-identifiable data. This trustworthy research environment will be linked to HPC through Microsoft Azure.
- Researchers can also use the ADRC-E's safe haven facility for secure data.

Training

- Relevant MSc programmes include: the health informatics component of the MSc in Health Leadership and management; an MSc in Data science; an MSc in Statistics with Applications in Medicine; and an MSc in Social Statistics; Other MSc programmes include: Artificial intelligence, Computer science, Cybersecurity, Economics and econometrics, Genomic Medicine, Geographical information systems, Health psychology, Official statistics, Operational research, Public health and Web sciences and MRes programmes in Clinical research and Advanced biological sciences.
- PhD programmes: The Faculty of Medicine hosts PhD students from the EPSRC-supported CDT in Web Sciences Innovation (15 pa.). The University of Southampton also holds relevant EPSRC Centres for Doctoral Training in Next Generation Computational Modelling and Complex Systems Simulation.

Public Involvement/Engagement

- The University of Southampton also target outreach and public participation in research, hosting LifeLab, Wessex Public Involvement Network and the NIHR's national INVOLVE hub.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Imperial College
- University of Oxford
- University of Dundee
- University of Edinburgh
- University of Leeds

NHS and other government organisations

- UHS Trust
- Southampton City Council; also, directors of public health in other councils across Wessex
- Ordnance Survey

Data providers

- CPRD
- NHS Digital

Industry

- Tech: Microsoft Research, My mHealth

International

- The EpiGen Research Consortium – partners include the Liggins Institute of the University of Auckland; and the National University of Singapore.
- Amsterdam
- Regensburg
- UBC Victoria

University of Surrey

Top-level insights

- The University of Surrey combines innovative engineering and technology with bioscience and has a strategic research focus on global wellbeing underpinned by innovations in health, and in particular, better use of digital technology to address health and healthcare challenges. To achieve this, it has established three informatics-relevant interdisciplinary themes: Lifelong Health and Wellbeing; Digital innovation; and Innovation for Health.
- One of the University of Surrey's key informatics assets is the Royal College of General Practitioners Research and Surveillance Centre analytical hub, which is the principal primary care research surveillance system for England that uses data from more than 250 general practices to monitor disease, measure vaccine effectiveness, and for research. Surrey also hosts the London Life Sciences Population (LOLIPOP) study health data (lead *Imperia*).
- Other investments in the last five years include, the 5G Innovation Centre (5GIC), which is the largest UK academic research centre dedicated to the next generation of mobile and wireless communications. It can offer a world leading, independent test bed for trialling emerging ideas, proving concepts, validating standards and vendor interoperability testing. The 5GIC is supported through a partnership of academia, government and industry, involving a large number (~60) of industry corporations and SMEs.
- Other collaborations with industry, but also NHS partners include the 'Technology Integrated Health Management for Dementia' test bed together with the Surrey and Borders Partnership NHS Foundation Trust. This is one of seven national innovation test beds; and one of only two focused on developing an Internet of Things for health.
- Surrey Research Park is a 70-acre University owned centre of excellence in technology, science, health and engineering. Adjacent to the partnership site, it provides a high quality working environment for over 110 companies including: BOC Linde, BAE Systems, IDBS, Microsoft, and EADS Astrium as well as over 100 SMEs and micro-companies. The Research Park data analytics businesses are University developed and support companies involved in the commercialisation of a wide range of sciences, including social science, technologies, health related activities and engineering.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The **5G Innovation Centre** (5GIC *Director: Professor Rahim Tafazoli*) is part of the University's Institute for Communication Systems, which sits in the Department of Electrical and Electronic Engineering. The Centre is supported through a partnership between academia, government and industry (£75M). It is the largest UK academic research centre dedicated to the next generation of mobile and wireless communications and offers a world-leading independent test bed for trialling emerging 5G ideas, proving concepts, validating standards and vendor interoperability testing.
- The **Engineering for Health** (£10M) is a new interdisciplinary hub that has as one of its aims to train the professionals driving technology for tomorrow's health service.

- The **Royal College of General Practitioners Research and Surveillance Centre** (RCGP RSC, *Director: Professor Simon de Lusignan*), established a secure data network and analytics hub within the University of Surrey's Department of Clinical and Experimental Medicine in 2015. For over 50 years, the RCGP's Research and Surveillance Centre has used data from more than 250 GP practices across England to monitor disease, measure vaccine effectiveness, and for research. The RCGP RSC is the principal primary care research surveillance system for England that includes linked virological sampling. *The scope of RCGP data are shown below:*

RCGP RSC RWE database:	Count as of Dec 2017
Data category:	
Current registered	2,026,150
Patients (ever registered)	4,947,304
Events (consultations)	1,351,807,943
Prescriptions	490,727,787
BMI Measurements	12,889,409
Blood Pressure Readings	30,594,700
Influenza vaccines given	6,049,683
Cases of influenza like illness	204,458

- Surrey's new **VetSchool** (£45M) leads new digital technologies in animal health (vHive) in partnership with Zoetis (£8.5M co-investment). These include: wearables, apps, sensors and satellites that promise to transform real-time information capture for animal owners and their veterinary advisors.

Knowledge/Expertise

- The University of Surrey has strengths in the rapid development of digital health and healthcare technologies and their application by providing a fertile technical and testing environment.
- It has robust expertise in the use of real world /routine data to monitor vaccine effectiveness and assess treatments including anti-microbial resistance; and develop health economic models.

Data

- The **RCGP Research and Surveillance Centre** uploads data from a nationally representative sample of >2M patients twice week. This is pseudonymised but linkable to other datasets. It is principally conducting surveillance on behalf of Public Health England, and vaccine effectiveness studies but also used in diabetes research. The database is nationally representative, and produces data in near real time, and has a dashboard capability that can provide insights into quality/morbidity to member practices.
- The **OPTIMAM2** project has developed a unique database of sequential mammography image data.
- Surrey's **Clinical Research Centre (Phase I MHRA accredited)** has a 14,000 healthy volunteer database, ready for invitation to trials. It conducts first in human (Phase I) trials. It also has state of the art sleep and chronobiology labs and data from a long history of this research is available for analysis. It networks with local clinicians and also works closely with the Surrey Clinical Trials Unit.

Translational and innovation capabilities

Links with industry

- The **5G Innovation Centre** (has provided £68M co-investment); involving 26 corporates and 40 SMEs with established arrangements for IP.
- Other commercial partnerships include pharmaceutical and animal health corporations involved in real world evidence studies and digital innovation. For example, Surrey's Clinical Research Centre works with leading pharmaceutical companies: Ono Pharma, Lundbeck, Merck, GSK, and Philips Lighting.

Links with NHS

- The University of Surrey works closely with **Royal Surrey County Hospital (RSCH)** NHS Trust. The partnership with the RSCH includes research collaborations, joint appointments, a framework agreement for clinical research services and aligned sponsorship processes. Developing this partnership is key to improving healthcare research, and they are working closely **Surrey Heartlands Sustainability and Transformation Partnership** attracting funding to increase the number of practices connected to RCGP RSC.. This is a five-year project (2017-2022), which aims to bring together clinicians and other health and care staff – alongside patients, their carers and families and members of the public – to think through how they can transform services so local residents have access to the very best care and treatments.
- Working in collaboration with **Surrey and Borders Partnership NHS Trust** and several other industry partners, the **Technology Integrated Health Management (TIHM) for Dementia** project, is based on an ecosystem of devices, which monitor a person's physical state and cues in their environment. TIHM is one of seven national innovation test beds; and one of only two focused on developing an Internet of Things for health. TIHM is funded by NHS England and Innovate UK.
- The University of Surrey is a founding member of the **Surrey Health Partners**, which includes a wider range of NHS collaborators (acute, mental health and community healthcare organisations) across Surrey. It also has strong regional links to the **Kent, Surrey and Sussex AHSN**.

Opportunities

- The University of Surrey offers robust knowledge and expertise in the development and testing of digital health interventions and a fertile environment to explore the next generation technologies in this area including the Internet of Things for Health.
- It has a track record of deep relationships with industrial partners but also NHS providers trying to address health and healthcare challenges through innovative technologies.

Potential gaps

- The University of Surrey does not seem to have significant activity in learning health systems research; although, it could develop this through the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) network.

Other relevant capabilities

Research themes

Please note that the University of Surrey Sol contains limited information on specific research theme activities. Some research activities of the PIs involved in the Sol submission can be found below:

Discovery Science

- application of computational biology approached to identify transcription-based biomarkers in circadian rhythms and cancer treatment (*Laing*);
- developing metabolomic and genetic models to understand how sleep deprivation interacts with circadian rhythms (*Skene, Dijk*);
- studies of genetic polymorphism to explore the nutrition impact on health (*Rayman*);
- early detection, stratification and treatment of cancer in male urinary tract cancer including the development of novel treatments (*Pandha*).

Precision medicine

- use of biomedical engineering approaches to identify better ways of delivering stratified medicine in neurological disease and sleep (*Abásolo*);
- use of advanced machine learning to find patterns of correlated symptoms in cancer treatment (*Barnaghi*);
- developing adaptive models to optimise cancer therapy, particularly in breast cancer (*Evans*);
- developing sophisticated models for improving the diagnostic harvest from a range of medical images, particularly in breast cancer; and in assessing kidney function (*Wells*).

Population and Public health

- influenza surveillance, diabetes, renovascular disease (*Lusignan*);
- use new technologies to detect outbreaks of zoonosis and making the food chain more efficient (*Cook*).

Citizen-driven health

- the **Remote Electronic Symptom Management system, eSMART** (including a €6M multi-national FP7 funded study entailing recruitment of patients undergoing chemotherapy across five countries);
- Food and Nutritional Science is very strong at Surrey including Queen's Anniversary Prize 2017, (*Lanham-New, Rayman, Griffin*).
- Online self-management of prostate cancer & digital solutions to support young adults with cancer.

Learning health systems

- There does not seem to be any major activity in this area based on the information provided in the Sol.

Disease areas

- Areas of strength are: sleep, circadian rhythms, and neurological disease; surveillance across human and veterinary health. Also, nutrition and its impact on health, diabetes, and cancer.

Analytics

- The Vet School's Bioinformatics core facility provides data analysis, mathematical modelling and computer simulation support. It has expertise in the analysis of genomics, transcriptomics and proteomics data.
- Other research activities include the development of: mathematical models for drug deposition in sleep (*Derks*); and health economic models based on routine data (*Gage*).

e-Infrastructure

- It also supports computational systems biology approaches, especially in the area of the constraint-based modelling of genome scale metabolic networks. It is increasing its 8 Linux servers comprising 2 - 64 CPUs, 1-128GB RAM and 16 -15000GB of storage.
- The activities of the Centre for Vision, Speech and Signal Processing (CVSSP) range across medical imaging, medical data analytics, machine learning, and big data. With over 1/3 of the university's total computer resource: 350TB storage, 700 processing cores + GPU processors facility equivalent of 1700 high end PCs.

Training

- In 2016, the University launched a new pioneering Doctoral College, aimed at both supporting the professional development of future research leaders (Post Graduate Researchers, PGRs and Early Career Researchers, ECRs) and building a cohesive community and research environment.
- Currently it offers a **short course in Clinical informatics** which plans to develop as a taught MSc in the future.
- It also offers Doctorate – PhD/MD – training in Digital health, machine learning and mHealth whilst plans are underway for PhD research projects with NHS using real-world data.

Public Involvement/Engagement

- The University of Surrey is working closely with the RSCH NHS Trust on the **Surrey Heartlands** project (see also NHS links section), which aims to bring together clinicians and other health and care staff – alongside patients, their carers and families and members of the public – to think through how they can transform services so local residents have access to the very best care and treatments.

Partnerships

- Overall, Surrey manages over 800 active partnerships with healthcare providers, government, corporate business, SME and the third sector to improve the delivery of health care nationally and internationally. These partnerships include: Macmillan Cancer Support, European Cancer Patient Coalition, CHASE Shooting Star Children's Hospice, VirginCare, Nuffield Health, North Downs Hospital, and overseas collaborations.

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- EPSRC; Innovate UK

NHS and other government organisations

- NHS Trusts: Royal Surrey County Hospital; Surrey and Borders Partnership; Surrey and Sussex; Surrey Heartlands STP
- Genomics England

Data providers

- Public Health England
- >250 General Practices

Industry

- Pharma: Merck, GSK, Zoetis, Lily, Takeda, Astra-Zeneca
- Tech: IBM, Philips

International

- FP7 eSMART project group, H2020 MOCHA, I-MOVE+ IMI ADVANCE, FLUCOP, DRIVE

GEOGRAPHIC REGION: NORTHERN IRELAND

- **Queen's University Belfast** submitted the only Statement of Interest (SoI) to become a Substantive HDR UK Site from **Northern Ireland (NI)**.

Cross-region Strategic Activities

- Northern Ireland has integrated health and social care within the NHS since 1974. This provides access to a rich source of well characterised data assets that can be effectively collated and mined. A crucial component of this data ecosystem has been the successful establishment and deployment of a fully implemented **Electronic Care Record (ECR)**. With a stable population of 1.8M, this integration makes NI a fertile environment to perform health and biomedical informatics research at the population level.
- The recent **eHealth and Care Strategy for NI** builds on the ECR's success and envisages the delivery of a fully integrated Electronic Health and Care Record system across the health and social care, providing complete and integrated patient information at point-of-care.
- The NI Department of Health and Social Care has a **Connected Health and Prosperity Agreement** with the Department for the Economy, underpinning a data-driven evidence-based strategy to enhance health and well-being of citizens through improved disease prevention and care, while also driving economic growth by enhancing competitiveness of NI industry. **A Connected Health Ecosystem** of healthcare, academic and business sectors has been established, focussed on addressing health and social care needs of the population.

Queen's University Belfast

Top-level insights

- Queen's University Belfast (QUB) capitalises on the advantage of Northern Ireland (NI) over other parts of the UK to have integrated health and social care within the NI NHS and a fully implemented Electronic Care Record (ECR). With a stable population of 1.8M, this integration makes NI a fertile environment to perform health and biomedical informatics research at the population level.
- In tandem, the Northern Ireland Cancer Registry, located on the QUB campus, was one of the first fully digital cancer registries in the world and crucially has 100% compliance in its capture of cancer incidence and mortality data.
- QUB researchers undertake nationally leading and internationally competitive research in digital pathology and image analysis, and have established industry partnerships in this area.
- QUB has internationally-leading expertise in data security and in the design and implementation of systems for large-scale data processing and retrieval supported with an HPC environment.
- QUB was not part of the Farr Institute but it has an ESRC Administrative Data Research Centre (ADRC-NI).

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments and Expertise

- The **QUB Bioinformatics Hub and the Informatics Core Technical Unit** provide academic leadership and core technical support to health informatics and bioinformatics related research.
- **QUB's Centre for Data Science and Scalable Computing (DSSC)** (*Nikolopoulos*) has research expertise in the design and implementation of software and system infrastructure for large-scale data processing and retrieval (with particular expertise in programming models, large-scale distributed data retrieval software and hardware technologies) to accelerate genome mapping, sequencing analysis, and processing of mass spectrometry metabolomics data.
- DSSC and the **Centre for Secure Information Technologies (CSIT)**; (£30M – EPSRC, IUK and NI), which is one of UK's lead university centre for cyber security research, are building together on their world-leading expertise in data security to develop solutions employing robust cryptographic algorithms that cannot be compromised by any future computing system, including Quantum computers. The recent award of a Regius Professorship, a rare and extremely prestigious award by her Majesty the Queen, highlights QUB's international reputation in this field (see more under 'Translational and innovation capabilities'.)

Data

- The **Northern Ireland Cancer Registry (NICR)**, located within QUB, collects, analyses and stores incidence and mortality data for all cancer patients in NI.
- The **ESRC Administrative Data Research Centre (ADRC-NI)** (*Director: Professor Dermot O'Reilly*) brings together QUB and Ulster University to provide access to de-identified administrative data routinely collected by government departments, and facilitate access to this data. Data includes: 100% census records, social security benefits data, screening, and health- related and environmental datasets.
- The **Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA)** involves the tracking of 8,500 subjects (>50 years), over a 10-year period, generating a comprehensive series of datasets that capture the dynamic social and economic relationships that shape health, social and personal wellbeing and economic position as people move into and progress beyond retirement.

Translational and innovation capabilities

- QUB's **Centre for Secure Information Technologies** (funded by EPSRC, Innovate UK and Invest Northern Ireland) is one of only seven UK **Innovation and Knowledge Centres**. These Centres are a key component of the UK's approach to the commercialisation of emerging technologies creating early stage critical mass in an area of disruptive technology. More specifically, the QUB Centre creates the security infrastructure needed to safeguard the trustworthiness of information stored electronically. As well as developing licensing deals and spin-out companies, it works closely with the entrepreneurial community to facilitate an environment that will foster innovation and entrepreneurship and ensure wider economic benefit.

Industry links

- QUB has a £3M programme with **Almac Diagnostics** that involves converting discovery science outputs into new biomarkers for clinical evaluation. Prognostic and predictive markers in colorectal and prostate cancer have been developed, permitting molecular stratification and development of new Precision Medicine trials in Oncology. There is significant collaboration between the Bioinformatics programme within Almac Diagnostics.
- **QUB – PHILIPS: Digital Pathology Partnership.** This partnership was born out of PHILIPS's acquisition of QUB spin out PathXL and the establishment of the PHILIPS Belfast Innovation Hub. The partnership focuses around Digital Pathology for molecular diagnostic workflow and tissue analytics, but will expand to develop an open ecosystem for Digital Pathology to drive transformation of tissue imaging and diagnostics.
- Partnerships have been established with major **hardware providers** (Intel, ARM, NVIDIA, Xilinx, Micron, IBM, Seagate) mainly through EPSRC, and Horizon2020, grants exploring innovative approaches to design, implementation and programming of micro-servers and storage systems for real-time data analytics at scale.

Opportunities

- QUB's world-leading expertise in data security to develop solutions employing robust cryptographic algorithms, and digital pathology and image analysis and integration with other data types can help catalyse HDR UK activities in these areas.
- The development of a fully integrated Electronic Health and Care Record system across the health and social care, providing complete and integrated patient information at point-of-care for a population of 1.8M will open a wealth of research opportunities.

- The recent Strategic Review of NI-NHS (**Bengoa Report**) and the subsequent Ministerial blueprint both focus on the need for a more stratified and person-centred approach to health. QUB is working in tandem with DoH-NI to support the establishment of the **Improvement Institute** flagged in the Minister's report, which places enhanced data analytics at the heart of efforts to improve health care delivery/public health and reduce health inequalities.
- An identified priority in the Ministerial blueprint is improving care for people with **dementia**. In this regard, a recent £7.4M award (*McGuinness*) will enhance data analytics infrastructure using the ECR for both discovery science and health care research for this disease.
- QUB's **Institute for Global Food Security** has a significant focus on nutrition and health. The Knowledge and Innovation Community award (£400M), secured through the European Institute of Innovation and Technology, with IGFS as a major partner, will generate an abundance of data that needs to be effectively mined in order to contribute to tackling the global challenge of nutrition, health and the feeding of the growing world population.

Potential gaps

- QUB does not seem to have significant research activity in citizen-driven/digital health and learning health systems.

Other relevant capabilities

Research themes

Discovery science

- QUB (*Salto-Tellez; Hamilton; James, McArt*) undertakes nationally leading and internationally competitive research in digital pathology and image analysis through the **UK Digital Pathology Network** (CR-UK Accelerator award £3.7M), this research, brings together academic and industry partners to drive a high-throughput, biomarker discovery and validation programme.

Precision Medicine

QUB leads or is a major partner in two **MRC Stratified Medicine Consortia**:

- **RASP-UK – Refractory Asthma Stratification Programme** (£4.8M; *Heaney*): assessing treatment adherence using remote monitoring technologies and employing biological markers to stratify asthma patients for therapeutic intervention.
- **S:CORT – Stratification in COloRectal Cancer Programme** (£5M; *Lawler, Johnston, Kennedy, Salto-Tellez*): this award, which is led by the University of Oxford (*Professor Tim Maughan*), aims to develop a clinical test(s) that could be given to all colorectal cancer patients upon diagnosis to enable tailor-made treatment plans, which is guided by their individual biological make-up.
- In **Chronic Kidney Disease** (CKD), an innovative Science Foundation Ireland-DfE funded project (*McKnight, Maxwell*) is integrating clinical, molecular and epidemiological data from >700,000 individuals to facilitate stratified medicine for CKD and deliver outcomes with economic and social impact.
- Additional partnerships in the precision medicine space (*Waugh, Gonzalez de Castro*), in particular with Almac Diagnostics (*Kennedy, Harkin*), contributing to the recent award of the **Precision Medicine Catapult** to Belfast as one of five UK Centres of Excellence.

Population and Public health

- The **Centre of Excellence for Public Health** (*Kee*) is one of five UKCRC-funded Centres in the UK. It builds on a platform of excellence in key programmatic areas including exploitation of existing health and social data, social determinants of health and disease and behaviours to promote and preserve healthy living. It hosts the Northern Ireland Cohort for the Longitudinal Study of Ageing (NICOLA), which underpins a cross-disciplinary research programme and allows detailed data comparative analysis with the English Longitudinal Study of Ageing (ELSA) and TILDA (The Irish Longitudinal Study of Ageing).

Citizen-driven/Digital health

- The **COMET initiative** is linking Patient Reported Outcome Measures (PROMs) to the comprehensive clinical trial, epidemiological, and 'omics datasets.

Disease areas

- QUB has research programmes on a number of disease areas with particular focus on cancer, kidney disease and asthma.

Analytics

- A key biomedical informatics output has been **PICan (Pathology Integromics in Cancer)**, a data analytics platform that captures diverse information sources (epidemiological, clinical, 'omics, images (TMAs, whole tissue sections, MRI), etc.) for rapid and accurate data mining and integration.
- The CR-UK Accelerator award has also led to the development and release of **QuPath**, a key open-source framework for performing digital imaging analysis.
- **TissueMark™**, a software product for automated identification/annotation of tumor tissue, developed in partnership with QUB spin-out company PathXL (*Hamilton*) received the Frost and Sullivan European New Product Innovation Award for Digital Pathology.
- Data scientists working with ADRC-NI have advanced novel linkage methodologies to preserve privacy and confidentiality of personal information.

e-Infrastructure

Compute

- QUB have established an HPC environment (*Kelvin*; £2M). The HPC cluster comprises 45 compute nodes running Linux and connected together with an InfiniBand fabric. Compute nodes range from 128GB to 1TB and are powered by Intel Haswell generation processors. Coupled to the HPC cluster is a 2PB usable multi-tier storage system, delivering a blend of scalable, high-performance, POSIX compatible file storage with a self-replicating object based filesystem. Kelvin currently has over 200 software applications available to users via a module files system.

Training

- QUB offers an **MSc in Bioinformatics and Computational Genomics** (*Blayney, McArt*) and an **MSc in Data Analytics** (*Marshall*), which is a new University initiative targeted specifically at building capacity in data science.
- A novel interdisciplinary Doctoral Training Programme (the £1M **Leverhulme Interdisciplinary Network on Cyber Security and Society**) supports pioneering research at the interface between social sciences, electronic engineering and computer science.

Public Involvement/Engagement

- The **Northern Ireland Cancer Research Consumer Forum** drives an agenda that emphasises the role of patients/citizens in research design and its implementation, both within NI and nationally. This Forum is embedded within S:CORT, contributing to the writing of the original grant submission, and bringing a PPI perspective to the research effort.
- Other activities in this space include partnerships between QUB (Lawler):
 - Welcome Trust Sanger Centre (*Middleton*) on the “**your-dna-your-say**”, a project exploring global attitudes around sharing genomic and clinical data)
 - Bonnie J Addario Foundation (*Valia*) on “**the Ros1ners**”, an online citizen-driven genomic data sharing project to accelerate precision healthcare.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- MRC Stratified Medicine Consortia: RASP-UK (QUB lead) and S:CORT (QUB partner, Oxford lead)
- ESRC ADRC-Northern Ireland
- UK Digital Molecular Pathology Network

Data providers

- NI Cancer Registry
- NI National Statistical Authority

Industry

- Technology companies: Philips, Intel, ARM, NVIDIA, Xilinx, Micron, IBM, Seagate
- SMEs: Almac Diagnostics

International

- GA4GH

GEOGRAPHIC REGION:

SCOTLAND

- Statements of Interest (Sols) to become a substantive HDR UK site were received from the following three ROs within the Scottish geographic cluster:

1. University of Edinburgh

2. University of Glasgow

3. University of Dundee

- Due to time/resource constraints, this report contains information only on ROs which expressed interests in becoming HDR UK Substantive Sites. The **Universities of Aberdeen, St Andrews and Strathclyde**, which are based in this geographic region, had also submitted Sols to become a Partner HDR UK but this report does not contain any information on their informatics/health data research activities.

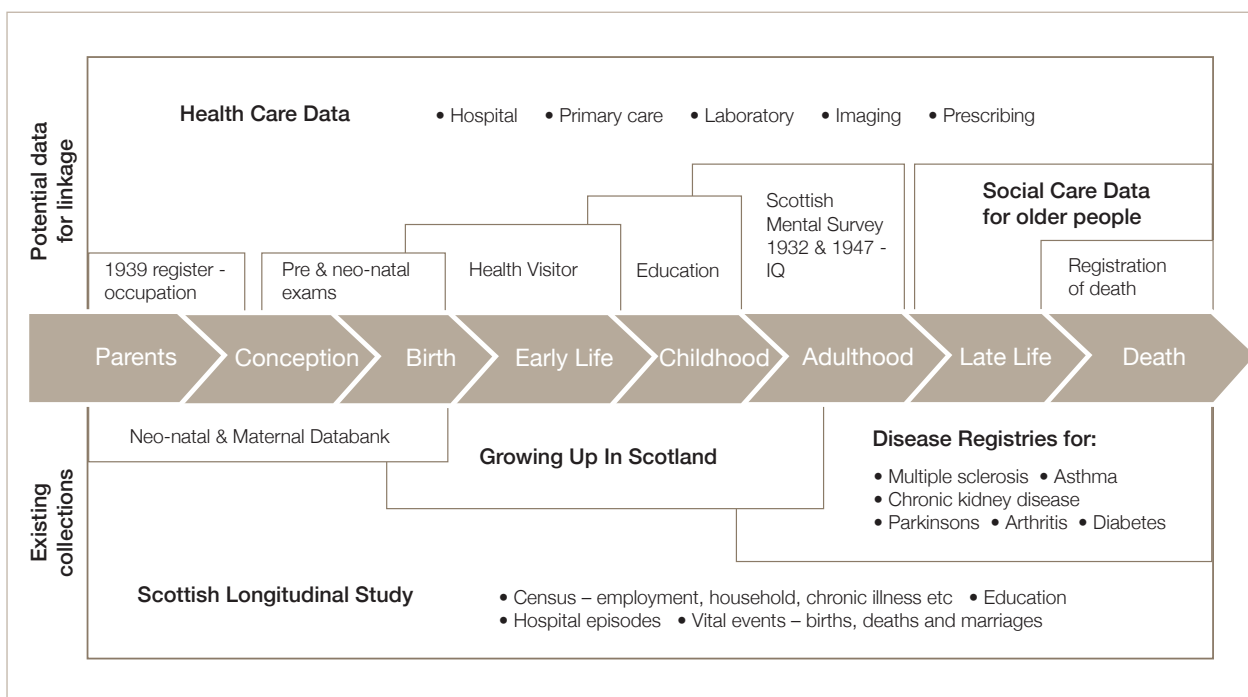
Cross-region Strategic Activities

- The Scottish Universities have a sustained track record of partnership working in health and biomedical informatics research. This is in collaboration with the Scottish Government and the NHS Scotland Territorial and National Boards including NHS National Services Scotland (NSS) that hosts the national datasets for 5.2M residents, dating back to 1980.
- Examples of large collaborative research investments in this area include:
 - **Farr@Scotland** (£13.5M) – please note that all six Scottish Universities that have submitted Sols were part of the Farr Institute in Scotland.
 - **ADRC Scotland** (£8M) – This is led by the University of Edinburgh in partnership with most universities in Scotland. The **Centre is co-located with Farr@Scotland**, and has developed the technical and physical infrastructure including safe settings in close collaboration with Farr.
 - **Generation Scotland** (£15M) is a record-linked bioresource is a unique partnership between the Scottish University Medical Schools, the NHS in Scotland and the people of Scotland, funded by the Scottish Government since 2003. Over 30,000 people from across Scotland have helped Generation Scotland create a world class biomedical resource for research into a wide variety of diseases. They did this by participating in a research study and contributing biological samples, clinical measurements and information about their health and lifestyle. Importantly, the majority have also given their permission to link to information from their medical records and to be re-contacted about participation in further research.
 - **SHARE** is a register of people (designed and hosted by the University of Dundee) interested in participating in health research and who agree to allow SHARE to use their various NHS EHR data to check whether they might be suitable for health research studies. SHARE is part of the NHS in Scotland and works in close partnership with researchers across Scotland. It currently has 147,000 (and rising) people consented for linkage of the NHS data to support direct recruitment, and includes consent to use spare blood from routine NHS testing.

- **Innovation Centres** (£120M) were launched in 2012 supported through the Scottish Funding Council and Scottish Enterprise and Highlands and Islands Enterprise. Of these, four are related more or less closely to **Big Data: Digital Health Institute** (based in Edinburgh and Glasgow); **Stratified Medicine Scotland Innovation Centre** (SMS-IC) (with headquarters in Glasgow); **CENSIS – Sensors and Imaging Systems** (based in Glasgow); and **DataLab** (based in Dundee, Edinburgh, Glasgow, and Aberdeen). The Innovation Centres have backing from industry and draw on all of Scotland's research expertise in the relevant sector to work on problems and opportunities identified by industry. They will add value through secondments, industrial studentships, spaces for collaborative work and shared access to equipment. The Centres will also support skills and training to develop the next generation of researchers and knowledge exchange practitioners through masters and post-doctoral level provision.
- These investments have helped create a Scottish health informatics ecosystem that is underpinned by:
 - a federated network of regional Trusted Research Environments (TREs)/Safe Havens located within Aberdeen (DaSH), Dundee (HIC), Edinburgh (HSRU) and Glasgow (NHS R&D); and
 - a national Safe Haven at National Services Scotland (eDRIS).

These TREs/Safe Havens operate independently to provide advice, support and a secure environment for access to a wide range of datasets (including national datasets through to specialised local datasets, see next bullet point) but also as a federated network across Scotland working to common principles and standards whilst common processes enable the safe and secure flow of data between them.

- Scotland has access to a wealth of national routinely collected health data and rich longitudinal cohorts, including but not limited to the ones presented in the diagram below:



- Furthermore, Scotland has a 30 year track record of nationwide record linkage research, using the Community Health Index (CHI) number for a population of ~5.2M, linked to a single health system for the entire country. This unique patient identifier is the key to linking health data for research purposes from all primary health care activities, and hospital-based clinical information systems in Scotland.

- Linked data includes: nationwide prescriptions (600M since 2009); the Scottish cancer registry (1.2M records since 1958); the Scottish Morbidity Record (52.5M hospital admissions since 1980); nationwide medical imaging data (21M examinations; 2009-2017); and a deeply phenotyped national diabetes dataset (4M patient years).
- In 2009, **Health Science Scotland** was formed to facilitate collaboration and promote excellence in the field of clinical and translational medicine. This **Academic Health Science Collaboration** involved four medical universities –Aberdeen, Dundee, Edinburgh and Glasgow- and their associated NHS Health Boards.
- From 2015, Health Science Scotland is part of **NHS Research Scotland** and brings together Government, Academia and the National Health Service, providing an operating structure that reaches across a wide range of disciplines, has access to all tiers of care, and can address virtually all of the chronic diseases that afflict developed countries.
- This close tripartite collaboration (Universities, NHS and Government) has spearheaded the development of health data policy, streamlining the information governance practices (single sign-off for Scottish projects; median 38 days).
- The Scottish Government and NHS Scotland have a national **eHealth Strategy** in place to support key health policy aims such as improving the quality of care, enabling shared decision-making with patients and integrating health and social care. Their vision for 2017 and beyond is person-centred and set out in the current **eHealth Strategy 2014-2017**. The strategy has a significant focus on healthcare and the needs of NHS Scotland. However, it is important to note that the strategic direction in Scotland is towards the provision of integrated health and social care, and this will be reflected in the future development of eHealth. Associated with this will be a further shift in focus from the internal needs of NHS Scotland towards joint requirements with local government and third sector partners, and, in particular, the expectations and requirements of citizens and patients for electronic information and digital services. In November 2016, they also announced the development of a new, integrated **Digital Health and Social Care Strategy (2017-22)**. A draft of this is currently under consultation.

University of Dundee

Top-level insights

- The University of Dundee has a long track record in health informatics research and has spearheaded developments in this area such as the Community Health Index now used as the unique patient identifier across NHS Scotland.
- It hosts a wealth of health data on over 1M individuals including one of the most comprehensive longitudinal dispensed prescribing datasets in the world, and 17% of the local population is consented into a bioresource.
- It also drives forward national initiatives such as SHARE, which has 200,057 recruited (and rising) people consented for linkage of their NHS data to support direct recruitment and for use of residual blood samples from routine NHS testing. It also offers the ideal platform for patient engagement, including capturing patient-reported outcomes and behaviour. The University of Dundee's ambition is to support SHARE implementation widely, and they are already deploying a local version in London.
- The University of Dundee has developed robust systems to provide tailored and rapid data linkage and data provision in their ISO27001 and nationally accredited Safe Haven.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The **Health Informatics Centre** (part of the Farr@Scotland) is a University of Dundee research support unit, within the Tayside Medical Science Centre, and works in collaboration with NHS Tayside and NHS Fife. It has over 35 staff and 15 years' experience in the secure hosting and handling of sensitive data and software development to support secure data collection.
- The **iCIGHT Centre** (informatics, Clinical Imaging, Genomics and Healthcare Technologies) is an interdisciplinary imaging research centre across four Schools (Medicine, Science and Engineering, Life Sciences, Nursing and Health Sciences) aimed to improve healthcare through computational and data analysis research.

Knowledge/Expertise

- The University of Dundee Health Informatics Centre is recognised as a leader in health data management and linkage underpinned by secure dataflows. Managing and analysing different types of imaging data at scale is an area of particular strength.
- It has also substantive expertise in developing patient recruitment services such as providing aggregate figures of study feasibility and potential study recruitment from various data sources.
- It also has extensive knowledge in the development and evaluation of decision support tools in clinical informatics systems and their implementation within the NHS as well as a track record of development of complex interventions to improve prescribing quality and safety.

Data

- The University of Dundee hosts [linked datasets](#) for ~1M people.
- Its [record-linked bioresource](#) includes more than 70,000 individuals (15% of Tayside population).
- As part of the **EMBARC project**, the University of Dundee currently runs large bronchiectasis registries for multiple countries, providing both independent national registries and the ability to securely analyse the federated dataset.
- The University of Dundee designed and hosts the **SHARE** research register which has 147,000 (and rising) people consented for linkage of their NHS data to support direct contact by researchers for recruitment, and includes consent to have spare blood from routine NHS testing. The University of Dundee Health Informatics Centre is currently implementing the SHARE system on behalf of a major partner in NW London.

Translational and innovation capabilities

Links with industry

- The University of Dundee has a growing relationship with NVIDIA and plans to collaboratively deploy advanced data analytics (deep learning) within its Safe Haven.
- Other major industry collaborations include the Astra Zeneca Global Genomics Initiative together with other Scottish Universities and NHS trusts.

Links with NHS

- As well as supporting interdisciplinary research within the University, the University of Dundee encourages collaboration between academics, industry and the NHS, for example through the establishment of the **Academic Health Science Partnership in Tayside (AHSP)** in 2014. Centred on a tripartite mission to deliver high quality clinical care, innovative healthcare education and training, and leading research and informatics, a core AHSP activity has been the development of informatics for research and innovation to drive care transformation.

Opportunities

- The University of Dundee's vision is to further develop the SHARE research register to create a shared bioresource of between 500K and 1M Scots linked to their EHRs. The SHARE infrastructure/approach, with population consent and engagement, and consent for collection of residual blood samples, could be potentially applied across HDR UK.

Potential gaps

- Currently, the activities of the Health Informatics Centre are not supported by substantial HPC resource to support deep-learning style big data analytics enabling discovery on large-scale data sets. However, they are trying to address this in collaboration with NVIDIA. The Centre is in the process of significantly upgrading their HPC resource over doubling the compute power, adding GPUs and storage capacity to support deep-learning style big data analytics enabling discovery on large-scale data sets.
- The University of Dundee does not seem to have significant activity in the area of digital health research.

Other relevant capabilities

Research themes

Discovery Science

- The **Computer Vision and Image Processing Group** (*Trucco, McKenna*), based in the Computing, School of Science & Engineering, has developed cutting-edge image analysis solutions for biomedical and assistive applications, including: histopathology, retinal imaging, MRI/CT, and video colonoscopy.
- The **Open Microscopy Environment** (OME) in the School of Life Sciences is a multi-site, collaborative effort involving both academia and industry which builds and provides open source interoperability tools for biological image data. OME applications were running in 4,603 locations and started 499,975 times in 2016.
- The **PepTracker** – [Encyclopedia of Proteome Dynamics Database & Computational Infrastructure](#) – provides a free to access, searchable, online database of large-scale, multi-dimensional proteomics data integrated with other large-scale ‘omics and genetic data sets.
- Dundee maintains two Elixir-UK Resources (Jalview, and the Dundee Resource for Protein Structure Prediction and Sequence Analysis) as part of the European research infrastructure in Bioinformatics
- University of Dundee researchers are involved in multiple genetics consortia including DIAGRAM, MAGIC, and GLC. In addition, the **Genomics Resources Consortium** has selected SHARE as one of its 11 international biobanks and is working with Dundee to pilot exemplar projects.

Precision medicine

The University of Dundee leads the:

- **IMI-DIRECT** (Diabetes REsearch on patient stratiFication) project (€46M) which aims to identify biomarkers that address current bottlenecks in diabetes drug development and develop a stratified medicines approach to the treatment of type 2 diabetes;
- data analytical strand of the MRC Stratified Medicine Consortium **MASTERMIND** (£6M);
- H2020 **PREDICTION-ADR** project (€6M) which aims to develop genetic risk assessment and diagnostic tools for the prediction of adverse drug reactions in two commonly prescribed drug classes that are used to manage cardiovascular risk, namely statins and ACE inhibitors/Angiotensin Receptor Blockers;
- bioinformatics analysis in the H2020 **ENSAT** project (€8.4M) which aims to develop and evaluate an ‘omics-based stratified health promotion programme for patients with endocrine forms of hypertension.
- Dundee leads the £7M NIHR Global Health Research Unit on Diabetes Outcomes Research focusing on stratification and risk prediction in Diabetes in India and Scotland
- EPSRC project (£1.1M) on identifying multi-modal retinal biomarkers for vascular dementia

Population and Public health

- The University of Dundee has an active programme of randomized trials of different recruitment strategies embedded in the studies it supports. The **Medicines Monitoring Unit (MEMO)** runs large, pragmatic, randomised trials using routinely collected data for follow-up to maximise efficiency and reduce costs of clinical trials compared with traditional industry-led studies. Currently, there are ~31,000 subjects randomised in MEMO studies.

- The University of Dundee also manages initial recruitment from primary care and other routine data of participants in research studies (for example, recruiting ~12,000 participants in 18 months to a recent cancer screening trial).

Citizen-driven/Digital health

- University of Dundee researchers have already created an app for SHARE participants where electronic and manually entered data is automatically integrated with data from health records, and will extend the scope of data collected in this way to include personal characteristics known to influence health, recovery and outcome.

Learning health systems

- University of Dundee works on the development and evaluation of informatics and routine data-driven interventions to improve healthcare quality and safety and has multiple completed, and in progress, prescribing improvement trials involving ~60% of Scottish general practices and every community dental practice in Scotland.
- The University of Dundee also works closely with the NHS to co-create novel informatics tools, for example providing *clinical leadership* to the **Infection Intelligence Platform** which provides integrated information on infection across Scotland by combining risk factor, demography, healthcare activity, medicines usage and clinical data.
- Other collaborative work with the NHS includes implementing decision support tools within national EHR systems such as **SCI-Diabetes** which is a good example of system support for long term conditions. This system is used by all diabetes services across Scotland and many general practices. It provides highly specialised functionality for the care of this specific condition, and is integrated into the wider eHealth infrastructure for information exchange. The **My Diabetes My Way** portal extends access to patients who sign in via myaccount, and a clinical decision support rules system is integrated to provide additional intelligence.
- Other work in this area includes the development and evaluation of system-wide informatics interventions to improve prescribing safety and to optimise medicines use in vulnerable people with polypharmacy.

Disease areas

- University of Dundee has research programmes in a number of disease areas with particular focus on: CVD and diabetes.

Analytics

- University of Dundee investigators have developed techniques for the computational analysis of biological sequences and three-dimensional structures. For example, **Jalview** is used by >70,000 sites each year to provide easy access to data on known DNA variants with genomic DNA, RNA and protein sequences and their three-dimensional structures in an evolutionary context.
- The **Computer Vision and Image Processing Group** works closely with the University of Dundee Health Informatics Centre and its **VAMPIRE** retinal image analysis software was the first home-grown image analysis package to run in the Centre's Safe Haven. The Group is partnering with NVIDIA to set up large-scale, deep-learning image and data analytics in the Safe Haven.

e-Infrastructure

Compute

- Underpinning the Health Informatics Centre services is a HPC virtual infrastructure running on both the NHS and the University networks with 812 cores, 6.44TB memory, 895TB storage, and 201 VMs.

Trusted Research Environment

- In the last five years, University of Dundee has invested more than £1M in the Centre to develop a state of the art software suite, the **Research Data Management Platform (RDMP)**, for the management, curation, anonymisation and rapid provision of research-ready data. The RDMP allows near real-time data provision to researchers from multiple linked data sources in an ISO27001 safe-haven environment, balancing strong information governance with rapid and efficient data provision. In the last 4 years the RDMP has been used to support over 500 projects, generating over 1300 data extracts of mainly phenotypic text based data for epidemiological research projects and clinical trials. It was the first Centre in Scotland to offer a Safe Haven, which is now Nationally Accredited.

Training

- University of Dundee hosts the MRC DTP in Quantitative and Interdisciplinary approaches to biomedical science (**QIBIOMED**), which has Big Data as one of its four focus areas.
- Also, it hosts a **Health Foundation Improvement Science PhD programme** in evaluation of NHS improvement using longitudinal data.
- The UoD **Masters in Public Health** programme includes modules on routine health data analysis and data visualization and a module in data analytics.
- In 2019, the University of Dundee will introduce an **MSc in Big Data and Precision Medicine**.

Public Involvement/Engagement

- In 2008, the University of Dundee developed **MyDiabetesMyWay** which gives patients access to their NHS health records integrated with home-recorded electronic and manually entered data such as home blood glucose measures, and uses sophisticated algorithms and data linkage to drive highly tailored self-management advice and reports.
- Through **SHARE**, the University of Dundee also provides a community platform for researchers to engage with patients and the public in research study design and service assessment, and have developed a SHARE mobile App to provide a secure and easy to use portal for participant recording of health and lifestyle data.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@Scotland partners include the Universities of Aberdeen, Dundee, Edinburgh, Glasgow, St Andrews, Strathclyde and Leicester
- EMBL-EBI

NHS and other government organisations

- NHS Research Scotland
- Academic Health Science Partnership in Tayside
- General Medical Council

Data providers

- Scottish Primary Care Research Network
- NHS National Services Scotland NHS Tayside, NHS Fife

Industry

- Pharma: AstraZeneca, Sanofi, Lilly, Boehringer Ingelheim, Novo Nordisk
- Tech: NVIDIA. Also OPTOS, Toshiba/Canon

International

- EU IMI and H2020 projects/consortia
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR

University of Edinburgh

Top-level insights

- Data science is viewed as a major strength and strategic opportunity across disciplines within the University of Edinburgh. For this reason its Sol to become a substantive site of HDR UK builds upon a considerable number of health and biomedical informatics research resources already established through a number of investments including: the University of Edinburgh Medical Informatics Centre; University of Edinburgh School of Informatics; the Farr Institute; and the Administrative Data Research Centre in Scotland.
- These investments are underpinned by the Edinburgh Data Science, a 1,000-researcher network, spanning the University maximizing the impact of the University's compute, data and analytics environment to forge new connections between disciplines, to deepen connections with the NHS and the public sector and to accelerate industry engagement.
- The Edinburgh Parallel Computing Centre (EPCC) hosts and manages: the UK's Supercomputer ARCHER; the DiRAC Blue Gene/Q system; and numerous smaller systems, including the EPCC Tier2 HPC, and the Edinburgh Compute Data Facility which also provides private cloud capability. Supercomputer services are available to both academic and industrial researchers.
- In addition to the significant HPC capability, other unique assets include a national federated network of trusted research environments (TREs) harbouring large structured and unstructured datasets of a population of 5.2M. These are supported by the development of a proportionate legal and governance framework for data sharing, embedded in health policy, allowing single ethics and privacy sign-off for Scotland.
- The University of Edinburgh hosts the Asthma Research UK Research Centre), which is the leading example of a disease-specific learning health system across the UK and provides a proven model to create disease-specific UK networks of multi-disciplinary research expertise.
- The University of Edinburgh has deep strategic partnerships with major data science technology providers and its computer science department has generated a number of start-ups/spin-outs (e.g. Edinburgh Molecular Imaging, Fios Genomics). It also has one of the largest research programmes in telehealth in the world (Telescot).
- This Sol has a major focus on opportunities in diverse areas of health and biomedical informatics research such as genome and imaging informatics, precision medicine, clinical trials and public engagement that could be capitalised by strengthening existing research capability and building on a wealth of existing assets available to the University of Edinburgh.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The **Usher Institute** of Population Health Sciences and Informatics sits within the Edinburgh Medical School, in the College of Medicine and Veterinary Medicine at the University of Edinburgh, and works collaboratively with partners across the NHS, government departments, industry and the public. It has three core centres in: Global Health Research; **Medical Informatics**; and Population Health Sciences. It also hosts: the **Edinburgh Clinical Trials Unit**; the Scottish Collaboration for Public Health Research and Policy; the Centre for Research on Families and Relationships; and the **Asthma UK Centre for Applied Research**. The Usher Institute also includes affiliated colleagues from the NHS as well as Scottish Government, with staff based within the **Farr Institute (Farr@Scotland)**, **ADRC-Scotland**, and **Edinburgh Data Science**.
- The **Centre for Medical Informatics** is newly established within the Usher Institute. The Centre has a unique joint leadership team which unites the fields of population health (*Sheikh*), data-intensive medicine (*Morris*) and computer science (*Robertson*). The mission of the Centre is to accelerate research into creating learning health systems and developing next-generation medicine through data-intensive informatics.
- The University of Edinburgh hosts **Farr@Scotland** (*Director: Morris*) together with Dundee University. This built upon a sustained track record of partnership working and trust in health and biomedical informatics research between the Universities of Aberdeen, Dundee, Edinburgh, Glasgow, Strathclyde and St Andrews. The **Farr@Scotland** is collocated with one of the four UK-based, **ESRC-funded Administrative Data Research Centres (ADRC-Scotland)**, which were established to make routinely collected administrative data, including social security, tax and education records, accessible for research. In Scotland, Farr and ADRC have integrated their governance and e-infrastructure to improve efficiency (thus achieving economies of scale, avoiding duplication) and now support over 300 academic, NHS and industry projects.
- The University of Edinburgh is a founding partner of the **Alan Turing Institute** (£80M; EPSRC) with Warwick, UCL, Oxford and Cambridge.
- In 2014 the University's priority was to establish the **Edinburgh Data Science** (EDS; convened by *Seckl, Robertson and Morris*). This cross-campus initiative brings together the Schools of Informatics, Mathematics, and Research Institutes and Centres to create a co-ordinated 1,000 researcher network of excellence that spans across the University from health and molecular biology, to science, engineering, social sciences, ethics and law.
- A key strength of the University of Edinburgh is its **School of Informatics** (*Moore*) with over 500 researchers. Other Institutes relevant to health data research include these in: Adaptive and Neural Computation; Computing Systems Architecture; Language, Cognition and Computation; and Perception, Action and Behaviour alongside the Laboratory for Foundations of Computer Science and Centre for Intelligent Systems & their Applications have created a powerhouse with extensive application to medicine.
- The **Edinburgh Robotarium** (centre for robotics and autonomous system facilities) is affiliated to the Usher Institute. This gives the University of Edinburgh a UK-leading capacity in robotic systems for use in hospital applications (e.g. prosthetics) or home settings (e.g. empathic systems for support of the elderly). This is complemented by the **Design Informatics Centre** which is a conduit to engagement with human factors specialists and product design specialists.

- The **Digital Curation Centre**, operated by the University of Edinburgh, is an internationally-recognised centre of expertise with a focus on building consultancy, capability, platforms, tools and skills for research data management.
- The University of Edinburgh is the lead partner for the **Software Sustainability Institute**, founded to support the UK's research software community, with a mission to cultivate better, more sustainable software to enable world-class research.

Knowledge/Expertise

- In addition to the significant experience of running supercomputer services, the University of Edinburgh has a long track record in data science and informatics including: artificial intelligence, advanced analytics, cybersecurity, automation of knowledge, cloud computing, data management, data linkage and disease modelling.
- Other expertise encompasses ethics, legal and social aspects of health data science.
- The University of Edinburgh has one of the largest critical masses of human molecular genetics and biology research capacity in the UK.

Compute

- The **Edinburgh Parallel Computing Centre** (EPCC; *Parsons*) hosts the UK's Supercomputer ARCHER and the research data facility (£96M government investment). Currently, it has 3,500 users accessing 118,080 cores and 28PB storage. EPCC works with the Farr Institute and the JISC Safeshare project to create a harmonised federated network of e-Infrastructure. Likewise, EPCC hosts the Scottish data platform for the Scottish hub of the ARDC.
- The Farr@Scotland at Edinburgh also hosts the **Scottish National Safe Haven and eData Research and Innovation Service**, which provides a single point of contact to assist in the completion of applications to the Public Benefit and Privacy Panel and assists researchers in study design, approvals and data access in a secure environment.

Data

- **Data assets curated at scale and managed locally** include the Scottish national PACS system (image data across Scotland at petabyte scale).
- In addition to the Scottish national datasets mentioned previously (*see also Geographic cluster: Scotland*), the University of Edinburgh also has an agreement with NHS Lothian to host regional and operational health and social care datasets for a population of 550,000 residents for research purposes.
- Furthermore, the University of Edinburgh has performed WGS on 3,500 people, permitting links to the longitudinal health informatics data held in Scotland, and other consented bioresources of >150,000 individuals including the UK Biobank (*Sudlow*), Lothian Birth Cohort (*Deary*), consented tissue banks in cancer (e.g. ovarian (*Gourley*), colon (*Dunlop*)), inflammatory bowel disease (*Satsangi*), Generation Scotland (*Porteous*) and human isolate studies in Orkney and Shetland (*Wilson*).
- The University of Edinburgh generates genomics data at scale. For example, the Edinburgh Genomics facility currently sequences ~200 whole human and animal genomes per week.

Translational and innovation capabilities

Links with industry

- The University of Edinburgh has established (MRC Proximity to Discovery; since 2014) an innovation programme in data-intensive healthcare, which placed them top in the UK for number of start-ups/spin-outs generated by a computer science department (according to SpinoutsUK). Spin-outs have included Edinburgh Molecular Imaging and Fios Genomics; and fostered two “unicorn” companies (exceeding \$1bn value).
- In addition, the University of Edinburgh has deep strategic partnerships with major data science technology providers (e.g., ARM, Google, IBM, Intel, Silicon Graphics, Microsoft, Selex) and consumers (e.g., Dixons Carphone, Genus, NHS Scotland, Pfizer, RBS, Aridhia, Standard Life Investments, Thermofisher, Zoetis); and relationships with a further 120 companies.
- The University of Edinburgh (*Morris*) led the Scotland/AZ collaboration including: the WGS and analysis of over 800 patients with ovarian cancer (*Gourley*); and the Biogen collaboration on MS (*Chandran*), both co-ordinated through the Stratified Medicines Scotland Innovation Centre.

Links with NHS

- The University of Edinburgh works in close partnership with NHS Scotland as evidenced by the large number of joint research programmes mentioned throughout this and other sections in this report.
- It is also a member of the **Academic Health Science Collaboration** together with three other Scottish medical universities: University of Aberdeen, University of Dundee and University Glasgow; and their associated NHS Health Boards.

Opportunities

By strengthening existing research capability and building on a wealth of existing assets, the University of Edinburgh has the potential to capitalise on opportunities that include:

- Exploiting advances in sequencing technology hosted by Edinburgh Genomics, including the Illumina X-10, Oxford Nanopore portable sequencing and the Illumina NovaSeq architecture, anticipated to enable the \$100-genome, to initiate a step change in the linkage of genomic/biological information and clinical data to develop medically relevant data infrastructures.
- Building on the nationwide diabetes programme (*Colhoun, Morris*), to create national deeply phenotyped cohorts and to expand to other chronic diseases. In cancer for example, work bring different stakeholders together with the Innovative Healthcare Delivery programme (£3M; *Keel, Cameron, Morris*) to develop a Cancer Intelligence Framework for the 220K people estimated to be living with cancer in Scotland providing near real-time tracking of every cancer case in Scotland from referral to assessment, molecular diagnosis and outcome. The same approach could be rolled out to other disease areas.
- Working on the challenging problems from health systems together with NHS Scotland and NHS Lothian to reconstruct consistent patient records across multiple media using machine learning, NLP and cognitive computing from actionable medical data in patient records that has little/no formal structure and is a mixture of text and images
- Creating a nationwide analytical platform for pharmacovigilance and phase IV “real world evidence” studies through the existing Scotland-wide partnerships.

- Addressing intelligent management, annotation and interpretation of biomedical images and applying their expertise in machine learning to provide a comprehensive and integrative approach to associate imaging data with other molecular and clinical data types. For example, develop deep learning platforms for automatically analysing large radiological imaging archives such as the 21M examination Scottish radiology archive. This would define and validate imaging biomarkers for cerebrovascular disease, dementia and other chronic diseases (*Wardlaw, Salman, Laurie*) or the >178,000 serial annual retinal images from people with diabetes to optimise the retinal screening programme and for predicting other vascular outcomes (*Colhoun, McKeigue, Storkey*).
- Building on the work of the Farr Institute to scale the proportionate governance and multi-strand public engagement work across the four nations and developing a viable social contract (*Laurie, Cunningham-Burley*).
- Working with the Alan Turing Institute, and other niche expertise in academia, NHS and industry to address key research challenges that mitigate against large-scale “big data” health research, challenges such as: automate data sharing; large scale social computation systems; real-time query language for exascale, federated genotype/phenotype data

Potential gaps

- The University of Edinburgh has limited research activity in citizen-driven/digital health except for telehealth.

Other relevant capabilities

Research themes

Discovery Science

- **Edinburgh Genomics** (*Aitman*) is a world leading genomics and bioinformatics facility delivering high volume genomics data and cutting-edge analyses to a large community of collaborators and customers across academia, government, and industry. It is one of the two genomics nodes of the NERC Biomolecular Analysis Facility, has National Capability core funding from BBSRC, and is one of three UK MRC High Throughput Sequencing Hubs.
- The **MRC Institute of Genetics and Molecular Medicine** (IGMM; *Frame*) is a strategic partnership of the:
 - **MRC Human Genetics Unit** (*Bickmore*) which aims to advance the understanding of genetic factors implicated in human disease and normal and abnormal development and physiology. We are part of the MRC Institute of Genetics & Molecular Medicine at the University of Edinburgh.
 - **Cancer Research UK Edinburgh Centre** (*Frame, Cameron*) whose clinical cancer informatics programme ambition is the deep phenotypic and molecular characterisation of the entire cancer population of Scotland.
 - **Centre for Genomic and Experimental Medicine** (*Aitman*) which aims to use genetics and genomics to understand the mechanisms of disease and improve disease prediction, prevention and prognosis.
- The **Roslin Institute** has a history in computational and mathematical biology, quantitative genetics and software development. It also has strategic investments such as the Illumina HiSeq X platform (£7.5M, *Watson*).
- In clinical imaging the **Edinburgh Imaging consortium** provides an umbrella for a broad range of activities (Brain Research Imaging Centre, Clinical Research Imaging Centre, retinal imaging and pre-clinical imaging facilities).

Precision medicine

- The **Scottish Genome Partnership programme** (£6M; MRC/CSO) is a major Scotland-wide research programme between the Universities of Edinburgh, Glasgow, Dundee and Aberdeen, NHS Scotland and NHS Lothian, NHS Greater Glasgow & Clyde, NHS Tayside and NHS Grampian. The research, which includes collaboration with Genomics England.
- The **Edinburgh-St Andrews Molecular Pathology Node** (*Aitman*) (£2M; MRC-EPSC) will bring molecular diagnostics into mainstream medicine by use of modern genome technologies and information across a range of diseases. The consortium will integrate state-of-the-art genomic and epigenomic methods for diagnosis of acutely ill children and will develop 'liquid biopsies' for managing cancer through analysis of circulating tumour DNA.

Population and Public health

- The **Edinburgh Clinical Trials Unit** embraces data science to transform the approach to Phase 2-4 clinical trials by supporting the design of intelligent trials using EHRs, by consolidating activities that use routinely collected data for RCTs, including cluster RCTs, and drug safety.

Citizen-driven/Digital health

- The University of Edinburgh has one of the largest research programmes in telehealth in the world (*McKinstry*). The **Telescot programme** (£14.6M since 2008) has conducted several large influential RCTs and many descriptive and qualitative studies in this area. These studies include telemonitoring of chronic obstructive airways disease, heart failure, high blood pressure and diabetes. Other studies of interest include: satellite location for wandering people with dementia, the use of gamification to encourage children and adolescents to adopt healthier lifestyles, developing artificially intelligent avatar based systems for monitoring depression, robotic dispensing and remote measurement of cough and respiratory rate.

Its current research efforts focus on the use of machine learning on patient accrued data to develop improved telemonitoring algorithms, video-consulting and the large-scale implementation of telehealth with Scale-Up BP. This project is set to change how general practitioners deliver blood pressure management thousands of people in Scotland.

- The University of Edinburgh has engaged in all the relevant Scottish Innovation Centres (DataLab, Stratified Medicine, Digital Healthcare, Sensors).

Learning health systems

- The University of Edinburgh hosts the **Asthma Research UK Research Centre** (£5M; *Sheikh; Griffiths QMUL*), which is the leading example of a disease-specific learning health system across the UK. This has created a UK-wide network of researchers who are working together to improve asthma control and reduce asthma hospital admissions and deaths. The Centre has created a: UK Asthma Observatory; a UK Forum for Patient and Public Involvement (PPI) in Asthma Research; a UK Database of Asthma Research Volunteers; and UK Methodology Service for Asthma Trials.

Disease areas

- Strong portfolio of disease-based programmes, specifically in cancer, diabetes, bone and joint diseases, gastro-intestinal diseases, psychiatric disorders, mental health and cognitive function, stroke and neurodegeneration.

Analytics

- Relevant activity was mentioned in other parts of this section.

e-Infrastructure

The University of Edinburgh has a broad range of Digital Research Services and e-infrastructures underpinned by a rich array of services and consultancy. These include:

- HPC and data management offered at layers of specialization and security appropriate to task. In particular, the University of Edinburgh **Information Service Group** provides a large suite of compute and data services focused around the **Edinburgh Compute and Data Facility (ECDF)** which provides general data storage (local or cloud) and compute services accessible from both users' workstations and large-scale computational services. ECDF has a layered approach with private cloud complemented by a well-provisioned 5,500 core compute cluster; a complementary General-Purpose GPU system; architectures that provide for large memory applications (using Symmetric Multi- processing with compute nodes up to 3Tb RAM); and 300TB of local storage with direct access to the University's 7PB DataStore.
- At the highest level of compute performance, the University of Edinburgh hosts and manages: the **ARCHER National HPC Service** (a 2.5PFlop Cray XC30) and **Research Data Facility** (23Pb) on behalf of EPSRC and NERC; the **DIRAC National HPC service** (a 1.6PFlop IBM BGQ) on behalf of STFC; and numerous smaller systems including: the new **EPCC Tier 2 HPC** and Research Data Facility; and a large shared memory system, funded by SFC, which supports health innovation and genomics projects. EPCC has invested in a proportion of each of these systems for general University use.
- The University provides direct support, data management and application consultancy for all groups and can provide specific support and advice for projects with secure data requirements (for instance currently supporting patient imaging data with University Hospital Research Institutes).

Training

- The **School of Informatics** offers a flexible choice of undergraduate and postgraduate courses and degrees in: Informatics; Artificial Intelligence; Cognitive Science; Computational Linguistics; Computer Science; and Software Engineering.
- In 2013, the University won share of a £350M investment in UK science and engineering postgraduate training. As part of this investment the School of Informatics now hosts two Centres for Doctoral Training and co-hosts a further centre along with Heriot Watt University. These include **Doctoral Training Programmes in: Precision Medicine** (*Saunders*); **Data Science** (*Williams*); **Robotics and Autonomous Systems, Pervasive Parallelism** (*O'Boyle*); and **Intelligent Sensing & Measurement** (*Underwood*). These are set to recruit >200 PhD students over the next three years.
- The University of Edinburgh has established data-science skills through doctoral and MSc programmes, the latter increasingly on-line, as well as through MOOCs and CPD short-courses.
- In addition, there are 90 tenure track Chancellor's **Fellows** appointed across the University strategically aligned to key themes notably genomics, medical informatics, data visualization, quantitative genetics, bioinformatics, imaging, ethics of data, and developmental or disease biology, supported by the **Edinburgh Scientific Academic Track** (*Bickmore*).

Public Involvement/Engagement

- The **JK Mason Institute for Medicine, Life Sciences and the Law** brings together legal, social and ethical expertise, in relation to health research regulation and data sharing (*Laurie, Cunningham-Burley*). Through this collaboration, the University of Edinburgh has built up substantial experience of conducting research-informed public engagement relating to the development of biobanks and data linkage for research, feeding into governance frameworks that have been adopted as best practice across all those domains.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@Scotland partners include the Universities of Aberdeen, Dundee, Edinburgh, Glasgow, St Andrews, Strathclyde and Leicester
- Farr@London (UCL), Farr@HeRC (Manchester) and Farr@CIPHER (Swansea)
- ARDC-Scotland and the other ADRCs and ADRN
- UK Biobank's Outcome Adjudication team (collaboration with Oxford, Swansea)
- DPUK
- Sanger Institute
- QMUL

NHS and other government organisations

- Scottish Government
- NHS Scotland, NHS Lothian, NHS Greater Glasgow & Clyde, NHS Tayside and NHS Grampian
- Genomics England
- Scottish Genome Partnership

Data providers

- NHS National Services Scotland
- Public Health Scotland

Industry

- Pharma: AstraZeneca, Pfizer, ThermoFisher
- Tech: ARM, Google, IBM, Intel, Silicon Graphics, Microsoft, Selex
- SMEs: Aridhia, Fios Genomics, Biogen

International

- Stanford (John Ioannidis)
- Harvard (David Bates, John Halamka)
- GA4GH
- ELIXIR UK – one of the 15 member organisations of the UK Node of ELIXIR
- Several IMI/H2020 informatics projects including: ROADMAP EPAD and RESCEU (Nair/Campbell)

University of Glasgow

Top-level insights

- The University of Glasgow has been one of the Farr Institute's partners, with the Farr@Scotland Deputy Director (*Pell*) and Farr's Training and Capacity Lead (*McCowan*) both based in Glasgow. Other collaborative programmes in health informatics include Generation Scotland and ADRC Scotland.
- The University of Glasgow is part of the Scottish Universities ecosystem with strong links to National Services Scotland and access to a wealth of national and regional datasets. For example, currently, the Glasgow Safe haven contains 40 datasets with more planned.
- The University of Glasgow played a leading role in developing the Scottish ecosystem for Precision Medicine, which brings together the four Scottish Universities with clinical medical schools, the corresponding NHS health boards and industry partners. The Stratified Medicine Scotland-Innovation Centre is led by the University of Glasgow and located in the Queen Elizabeth University Hospital campus, which houses one of Europe's largest hospitals.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- The University of Glasgow is partner in the Farr Institute's Centre in Scotland (**Farr@Scotland**) and **ADRC Scotland**.

Knowledge/Expertise

- The University of Glasgow researchers have strong expertise in the translation of research into healthcare, and in particular precision medicine, demonstrated by their involvement in a number of national and international programmes together with the NHS and industry partners. For example, an overall investment of more than £30M is integrating research with the NHS to transform current treatment approaches to pancreatic cancer; and a similar approach is used for ovarian cancer and COPD.
- In addition to precision medicine, the University of Glasgow researchers have a long track record in the development of innovative methodology for clinical trials and in particular, changing the approach to clinical trial/study participant follow-up.

Data

- The **Glasgow Safe Haven** currently contains 40 datasets with more planned. Around half are national datasets and half are specific to the West of Scotland including deeply phenotyped disease registries (rheumatology, Parkinson's, psychosis and cancer); intervention databases (weight management programmes and cardiovascular primary prevention programmes); and social care data on vulnerable groups. A full list of data resources available to UoG is provided in their Sol's Annex C.
- The **Glasgow Biorepository** is a joint academic-NHS venture. The NHS obtains electronic consent from patients for the storage of surplus clinical (blood and tissue samples). These are linked to phenotypic data in the Glasgow Safe Haven enabling selection of appropriate samples for research.

- The **Scottish Veteran Cohort Study** is a virtual cohort derived entirely from routinely collected health data. Primary care records were used to identify every Scottish resident who had entered and left military service. The 57,000 veterans, and a matched comparison group of 171,000 non-veterans, were linked to their acute and psychiatric hospital admissions, cancer registrations and death records to provide comprehensive follow-up data.
- Glasgow local authority partners have contributed several social care databases in the Glasgow Safe Haven for linkage to health datasets. Through a partnership with the Department of Education they have achieved the first Scotland-wide linkage of health and education data.
- The **Centre for Research on Environment Society and Health** is a virtual centre across Glasgow and Edinburgh Universities. The Centre uses linkage of health, environmental and social data to identify harmful and positive influences on health, with a particular focus on the natural and built environment.
- It has access to the **Urban Big Data Centre** that brings together urban social scientists, data scientists and statisticians to undertake cutting-edge research using data on the urban environment whilst the **MRC/CSO Social and Public Health Sciences Unit** is currently developing a longitudinal environmental dataset that can be linked to existing longitudinal health data.

Translational and innovation capabilities

Links with industry

- The University of Glasgow has established a **Clinical Innovation Zone** in their new Queen Elizabeth University Hospital campus where academia, industry and NHS are co-located to provide a precision medicine ecosystem. This includes:
 - the Stratified Medicine Scotland Innovation Centre (SMS-IC)
 - **Aridhia** – a world-leading clinical and translational informatics company, developing technology and capability that helps transform clinical research into clinical practice
 - **Clinnovate Health** – a Singapore and UK-based clinical service and research organisation focused on integrating clinical and research expertise with health innovation through novel diagnostic platforms
- Diagnostic images linked to other healthcare data, accessed via the Glasgow Safe Haven, to undertake innovative research are used in collaborative initial projects with **Toshiba** (brain CT) and **Butterfly Inc** (echo) to support testing, validation and optimisation of diagnostic characteristics (detection and quantification) through machine learning approaches.
- A novel platform *currently* being piloted called **CHART-ADAPT** aims to connect healthcare and research through data analysis provisioning technology. This is a collaboration between the University of Glasgow, NHS GGC, Aridhia Informatics and **Philips Healthcare**. The platform will enable important physiological models to be developed and validated using clinical data collected in the Intensive Care Unit, with results returned immediately to the bedside for use by NHS clinicians.
- The University of Glasgow is one of five UK partners of the **EU IMI EHR4CR** (16m euros – one of the largest **IMI** projects) project aiming to transform efficiency and speed of large, multi-national clinical trials by developing a platform for testing trial feasibility and identifying and recruiting eligible participants via distributed querying of multiple systems located across multiple sites/countries. This includes acting as a EHR4CR Champion Hospital using its commercial product/platform to identify and recruit eligible trial participants from across multiple sites/countries.

Links with NHS

- The University of Glasgow close working relationship with the NHS is exemplified by their joint partnership in a number of initiatives including: Glasgow Safe Haven; Glasgow Tissue Repository; SMS-IC; Imaging Centre of Excellence; EHR4CR; and CHART-ADAPT.

Opportunities

- A new campus development will include a newly constructed building for the multidisciplinary Institute of Health and Wellbeing; one of the Institute's three Research Themes is Data Science.

Potential gaps

- University of Glasgow does not seem to be leading on any important research programmes on novel deep data analytics. Instead it depends on collaborations with industry to provide data analytics solutions.
- It also does not have significant research activity in citizen-driven/digital health.

Other relevant capabilities

Research themes

Discovery Science

- **Glasgow Polyomics** is a Glasgow University facility that applies state of the art technologies to measure the genome, transcriptome, proteome, and metabolome from any biological system with a view to obtaining new biological inference by combining multi-level polyomics datasets.
- The **Imaging Centre of Excellence** (£32M; MRC, UK Research Partnership Investment Fund and European Regional Development). The facility is linked to the hospital providing direct access to clinical patients. It houses state of the art imaging technology including the only 7T MRI scanner in Scotland.

Precision medicine

- **Precision Panc** is a dynamic therapeutic development platform (led by Glasgow – *Biankin*) that enables precision oncology across the UK. This >£30M overall investment from numerous sectors, including industry partners, is integrating research with the NHS to transform current treatment approaches to pancreatic cancer.
- **AIM HY** is an MRC/BHF funded stratified medicine consortium involving 26 investigators across 11 academic institutions in the UK and USA. The consortium is investigating whether genetic, epigenetic and biochemical markers can predict the most effective drug or drug combination for the treatment of hypertension. The University of Glasgow is leading the metabolomics workstream and is a key recruitment site.
- The University of Glasgow plays a leadership role in other precision medicine programmes in ovarian cancer, COPD, MS and rheumatoid arthritis.
- The **Glasgow Molecular Pathology Node** (£3.4M; MRC-EPSRC) integrates pathology, genomics, and informatics and brings academics, clinicians and industry together to develop molecular diagnostic tools for use in disease stratification.

Population and Public health

- The **Glasgow Centre for Population Health** is jointly funded by Glasgow, NHS and Local Authority with the aim of championing joint working on public health priorities.
- The University of Glasgow researchers (*McCowan, Brittenden*) use aggregated healthcare data housed in their safe haven to test the feasibility of clinical trials and then use individual data to identify and contact eligible participants.
- In addition, University of Glasgow researchers have validated five-year follow-up using linkage to routine data against the gold standard methodology, and recently used this approach to evaluate efficacy over 20-year follow-up in the West of Scotland Coronary Prevention Study.

Citizen-driven/Digital health

- It does not appear that the University of Glasgow has any noteworthy activity in this area.

Learning health systems

- Data on all patients attending the Glasgow Hospitals with acute coronary syndrome are now accessible for research within the safe haven with findings from research fed back directly into the NHS reporting systems to improve clinical management.

Disease areas

- The University of Glasgow has research programmes in a number of disease areas including: cancer, COPD, CVS, MS and rheumatoid arthritis.

Analytics

- University of Glasgow researchers (*Hawkins*) specialises in the health economics of precision medicine; using Bayesian approaches to health technology assessment to prioritise precision medicine research and inform study design.

e-Infrastructure

Compute

- The new campus development has plans to include a Data Centre which will provide a secure data storage facility for clinical trials, epidemiological studies and the Safe Haven. The facility will meet ISO 27001 Information Security Management accreditation.

Trusted Research Environment

- Glasgow Safe Haven – has devolved research ethics committee approving power. It currently contains 40 datasets with more planned.

Training

- **Collin McCowan** leads the **Farr Institute's Capacity Building** programme and instigated the ambitious **Farr Future Leader Programme** which is developing an annual cohort of 12 competitively selected, international, mid-career researchers into the next generation of leaders in Health Informatics.
- The University of Glasgow runs seven directly relevant **MScs** in: Bioinformatics; Polyomics & Systems Biology; Public Health; Clinical Trials & Stratified Medicine; Stratified Medicine & Pharmacological Innovation; Health Technology Assessment; and Medical Genetics & Genomics, with 140 students currently enrolled in these courses.

- Glasgow and Edinburgh Universities also offer a joint **MRC Doctoral Training Programme in Precision Medicine** which offers students a 3.5/4 year PhD studentship award in health/biomedical informatics.

Public Involvement/Engagement

- The University of Glasgow has established disease specific **Patient Involvement Groups** and set up peer research groups to enable general public representatives to be partners in co-produced research.
- The University of Glasgow are partners in the **SHARE**, NHS Research Scotland initiative.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@Scotland partners include the Universities of Aberdeen, Dundee, Edinburgh, Glasgow, St Andrews, Strathclyde and Leicester
- Other Scottish Universities: Highlands and Islands, and Stirling

NHS and other government organisations

- NHS Research Scotland
- Genomics England
- Scottish Genome Partnership

Industry

- Pharma/diagnostic: AstraZeneca, ThermoFisher Scientific, Illumina, Clinnovate Health
- Tech: Aridhia, Toshiba, Butterfly Inc, Philips Healthcare

International

- EU IMI EHR4CR

GEOGRAPHIC REGION: WALES

- Statements of Interest (Sols) to become a substantive HDR UK site were received from two Research Organisations within the Welsh geographic region:
 - **Cardiff University; and**
 - **Swansea University.**

Cross-region Strategic Activities

The Government's agenda for this region

The UK and Welsh Government are investing over £2B in two City-region deals in Swansea and Cardiff, that over the next 20 years, and will see the generation over 35,000 new jobs.

- The [Cardiff Capital Region Deal](#) brings £1.2B and will see £495M invested including the creation of an “innovation district”, as well as supporting projects around data innovation, software development and cyber security.
- The biggest ever UK and Welsh Government investment for south west Wales (£1.3B) has been secured following the approval of the **Swansea Bay City Deal** that will transform the economic landscape of the area, boost the local economy by £1.8B, and generate almost 10,000 new jobs over the next 15 years. Eleven major projects will get underway, delivering world-class facilities in the fields of energy, smart manufacturing, innovation and life science, with major investment in the region's digital infrastructure and workforce skills and talent underpinning each.

Region-wide health and biomedical informatics activities

- In their **Informed Health and Care Strategy** the Welsh Government sets out its desire for health and care workers to have the digital tools and access to information they need to coordinate and deliver care, and for individuals to be empowered to play an active role in decisions about their own health and wellbeing through access to their own health and care information.
- The **Taking Wales Forward**, the Welsh Government's 2016-2021 strategy places a focus on putting health at the heart of everything they do and includes their ambition to exploit digital technologies to help speed up the diagnosis of illness. This includes making better use of data including improving data collection and quality; developing a more transparent framework for how it is managed and shared; and developing the skills and resources to realise its full value.
- As part of the Government's Swansea Bay City Deal, the [ARCH consortium \(a Regional Collaboration for Health\)](#) plans to enable an enhanced multi-sectoral data linkage on a population of 1M in South West Wales. This will drive the case for change in modernising health services, including the development and evaluation of a Wellbeing Programme. The ARCH partners are working together to further enhance the reputation of South West Wales, and indeed the whole of Wales, as a centre of excellence in service delivery, research, education and innovation in terms of healthcare and life sciences.

- The **'Shaping our future wellbeing strategy 2015-2025'** which is Cardiff and Vale University Health Board's vision of care over 10 years. The health board is working in partnership with Cardiff University, the City of Cardiff Council and the Vale of Glamorgan Council as [the Health Enterprise Alliance for Regional Transformation](#) (HEART), creating an ambitious vision for innovation, integration and citizen focused health and wellbeing in the region.
- Both the University of Cardiff and Swansea have access to considerable e-infrastructure spanning HPC and cloud computing through their membership in two consortia: Supercomputing Wales (SPW) and MRC Medical Bioinformatics CLIMB that provide researchers from across UK academia with access to computational infrastructure. All systems are co-located in the Redwood Data Center, and processes and policies are either already in place or in development for data transfer between systems. With regards to SPW and CLIMB, Cardiff provide the technical leadership for both projects, and is also the regional lead for CLIMB.

The **Supercomputing Wales (SPW)** consortium is five-year programme (2017-2022) is funded by the European Regional Development Fund (£9M) through the Welsh Government. This pan-Wales distributed supercomputing network is one of the largest in the UK – with eight sites across the country based on a hub and spoke model with hubs in Swansea and Cardiff – and allows companies and academics to access its 17,000 processor cores remotely. With the support of peak processing performance of almost 320 TFlops, businesses and academics can use the technology to power breakthroughs in product, process or service development; strengthening their market position. The project has been designed to support the public and private sector to deliver world-class projects in collaboration, removing the barriers to innovation and delivering sustainable economic growth and job creation.

- The [All Wales Medical Genetics Service \(AWMGS\)](#) is Wales' nationwide service for providing specialist genetic services for rare genetic conditions and cancers. It provides consultant led clinics for rare inherited disorders in hospitals in all health boards and also provides specialist and multidisciplinary clinics at University Hospitals, including for specific disorders/groups of disorders in which Welsh researchers have major interests. Specialising in healthcare service delivery, AWMGS also works closely with academic preclinical research through partnership with Wales Gene Park with which it shares next generation sequencing facilities. As part of Welsh Government's Genomics for Precision Medicine Strategy (>£6 million) AWMGS is currently expanding its laboratory and bioinformatics workforce, sequencing and IT infrastructure, the latter through collaboration with Wales Gene Park, Cardiff University, and Supercomputing Wales.
- The Welsh Government **'Genomics for Precision Medicine strategy'** was launched in July 2017, which saw the formation of two dedicated units – the Pathogen Genomics Unit within Public Health Wales NHS Trust and The All Wales Genetics Service within Cardiff and Vale NHS Trust – which will provide clinical genomics services for the whole of Wales. These activities are overseen by a national Genomics Taskforce, and underpinned by multi-million pound investment by Welsh Government. Both of these units are closely coupled to Cardiff University, and are principally based at the University Hospital of Wales site, alongside Cardiff University's School of Medicine. In both cases, the genomics work undertaken within the NHS is increasingly underpinned by computational resources at Cardiff University, to provide value for money and create opportunities for NHS data sharing and reuse.

- **Wales Genomics Medicine Centre (GMC)** – The Wales Genomics Medicine Centre (GMC) is Wales contribution to Genomics England's 100,000 Genomes Project. Awarded in July 2017 and formally launched on 25th January 2018, GMC joins UK's 15 existing GMCs. The Wales GMC is being delivered by the All Wales Medical Genetics Services (AWMGS, hosted by Cardiff & Vale UHB), and financially supported by Cardiff University, through MRC funding, in collaboration with the Wales Gene Park and the Welsh Government. The partnership is an important opportunity for patients to access potential clinical benefits via the 100,000 Genomes Project, and for Wales to contribute to the Project's research and other aims around health system transformation. As part of the MRC award, the **Genetic Data Integration (GeDI) scoping study** has been established by the [Wales GMC](#) with partners, the SAIL (Secure Anonymised Information Linkage) Databank, to study the data governance issues around integrating genetic data with health-related records for research.

Cardiff University

Top-level insights

- Cardiff University's strengths in health data science focus on genetics/genomics and population research, but currently its activities in this area have expanded to include imaging and metabolomics/ lipidomics.
- Its strong partnerships with NHS across Wales (population of 3.5M) and other government departments has enabled Cardiff University to link: research data to routinely collected health and social care data creating unique data assets; and provide a test-bed for the delivery of informatics-driven improvements to clinical services and public health interventions.
- Its research strengths are underpinned by investments over the past 10 years in HPC facilities, with further multi-million pound investments due to come during 2017-18 as part of Supercomputing Wales and GW4 Tier-2 HPC Centre developments.
- Cardiff University is one of the Farr@CIPHER partners with main contributions on the analysis of the social, environmental and biological determinants of health, development and education outcomes during childhood, and epidemiology of infections.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- Cardiff University's **Data Innovation Research Institute** (£2M; University investment) has been set up to conduct fundamental research into the aspects of managing, analysing and interpreting massive volumes of textual and numerical information. In the biological and life sciences, the Institute's research predominantly focuses on tackling some key challenges including: extracting information from data sets without compromising privacy and confidentiality; interpreting large data sets into reliable and understandable mathematical models; genome mapping; MRI scanning; drug trialling; maintaining public health records; and modelling protein structures.
- As part of the [MRC Medical Bioinformatics investment into the Cloud Infrastructure for Microbial Bioinformatics](#) (CLIMB; £8.5M), Cardiff is one of four distributed sites (Warwick, Birmingham, Swansea and Cardiff) and has provided the technical leadership for the national system, as well as acting as the regional lead for CLIMB in Wales. CLIMB provides infrastructure based on OpenStack, with a variety of virtual server infrastructure provided to the research community. In addition to providing traditional data storage and compute infrastructure, this provides a large geo-replicated object store both to support longer-term storage and to underpin collaboration. The project is working on the automated provisioning of genomic software and pipelines to assist the microbial genomics community. CLIMB currently supports over 1,000 microbiological researchers across the UK (*see also section 2.1.2*).
- Cardiff University has announced the establishment of a **Centre of Excellence in Cyber Security Analytics** that will be located at the School of Computer Science and Informatics and will be the first centre of its kind in Europe. Together with experts from Airbus, researchers will carry out world-leading studies into machine learning, data analytics, and artificial intelligence for cyber-attack detection. This research will aim to protect corporate IT networks, intellectual property, and critical national infrastructure and attempt to fill the skills gap that currently exists in the field.

Knowledge/Expertise

- Cardiff University has long-standing know-how in large-scale epidemiological studies using record-linked health and education data underpinned by appropriate data governance frameworks.
- There are also data governance and cybersecurity expertise within Computer Science and the School of Law including over a decade of experience of working with the NHS to enable data sharing between the NHS and University (including the provision of physical infrastructure for secure data transfer, systems design to enable data to be moved in/out of the NHS network, and policy development to support these activities and ensure they meet NHS governance requirements).
- Cardiff University has expertise gained in designing and providing the technical leadership of national e-infrastructures (Supercomputing Wales and (MRC CLIMB)); and for analysing biological data using cloud computing.

Data

- Cardiff is leading **HealthWise Wales**, Health and Care Research Wales' flagship national population cohort study that will provide the platform for precision medicine research via a well-characterised study population with longitudinal follow-up data and record-linkage to routinely available healthcare data. This project is on track to recruit 260,000 adults age 16 years or above by 2020, with planned enhancements for bio-sampling, cognitive assessments and recruitment of children and families, that will enable the investigation and follow-up of younger cohorts (not currently available through other population cohorts such as UK Biobank/Million Women Study) to inform strategies for the prevention of non-communicable diseases.
- The **Centre for Trials Research** has established frameworks for linking trial datasets with routine data held by NHS, Departments of Health (e.g. abortions data) and Education (e.g. National Pupil Database). Furthermore, it has over 75 current active studies that include more than 20 thousand people with associated clinical samples deposited into **Biobank Wales**.
- The **Brain Research Imaging Centre** (CUBRIC) is amongst the largest in Europe and hosts the National Facility for Microstructural Imaging. In addition, it collects medical imaging data and linked to genetic data.
- The **MRC Centre for Neuropsychiatric Genetics and Genomics** has access to very large genomic datasets of psychiatric, neurodevelopmental and neurodegenerative disorders (over 70,000 exome sequences, along with genome-wide association data on over 100,000 individuals) together with detailed phenotypic data.
- The **National Centre for Mental Health** is a Wales-wide initiative linking genetic and detailed phenotypic data from individuals with mental health disorders with routinely collected health data via the SAIL Databank, with the potential to yield genetically and clinically meaningful categorisations that go beyond traditional diagnostic criteria.
- The ERC/Wellcome Trust funded **Cardiff Lipidomic Group** hosts LIPID MAPS (£1.4M), a global online database with lipidomics/metabolomics mass spectrometry data.

Compute

- Cardiff University has considerable e-infrastructure spanning HPC and cloud computing. In addition to local infrastructure, CU is a member of three consortia: MRC CLIMB; Supercomputing Wales, and GW4 Tier-2 HPC Centre that provide researchers from across UK academia with access to computational infrastructure hosted at Cardiff. All systems are co-located in the Redwood Datacenter, and processes and policies are either already in place or in development for data transfer between systems.
- The **GW4 Alliance**, brings together four of universities in the south west of the UK; the universities of Bath, Bristol, Cardiff and Exeter. Together with Cray Inc. and the Met Office, GW4 Alliance has been awarded £3M by EPSRC to be installed in 2017 and deliver a new Tier 2 HPC service for UK-based scientists. The system will be one of the world's first to be based on Broadcom's Vulcan server-class chip. Details of this device are still under NDA, but the Vulcan CPU is generating excitement because it trades off much greater provision of memory bandwidth for less emphasis on peak FLOP/s, the former being more important for most scientific codes. Providing access to such a machine as a national service should therefore enable the UK's HPC community to quantify the benefit of memory bandwidth focused CPUs, thus informing future system procurements from Tier 1 to Tier 3.

Translational and innovation capabilities

Links with NHS

- Cardiff is leading on the newly established **Wales Genomic Medicine Centre** (a collaboration between Cardiff University, NHS Trusts, and SAIL Databank (£1M MRC) to work with Genomics England).
- The Cardiff and Vale University Health Board hosts the **NHS All Wales Medical Genetics service** and benefits from a close liaison with medical genetics research laboratories at Cardiff University due to co-location with the Welsh Government funded [Wales Gene Park Genomic Facility](#).
- Cardiff University provides the bioinformatics leadership for the newly established **Pathogen Genomics Unit** within Public Health Wales NHS Trust (with £1.2M Welsh Government funding), which will be providing Pathogen genomics diagnostics and epidemiological surveillance for the whole of Wales.

Opportunities

- The combination of CU's health data assets and strong computational infrastructure together with the opportunities offered by the NHS in Wales to deploy approaches across a healthcare system at a national scale provides a great opportunity to utilise Wales' population of 3.5M and devolved healthcare system as test-bed for health data research.

Potential gaps

- Cardiff University does not seem to be leading on any important research programmes in citizen-driven/digital health and learning health systems.
- It also does not seem to have strong collaborations with the health data/technology industry.

Other relevant capabilities

Research themes

Discovery Science

- The Cardiff University-led **Wales Gene Park** (£2.5M Welsh Government) takes advantage of the available HPC facilities to advance its cutting-edge research that provides understanding, diagnosis and treatment of a wide-range of inherited diseases and cancer. Furthermore, it has independent computational resources along with sequencing capacity, and is aligned with the **NHS All Wales Medical Genetics Service**.
- The **Cardiff Lipidomic Group** uses mass spectrometry to discover new lipid mediators of innate immunity and inflammation. In addition, they have also received European Research Council funding (€2.9M, 2014-2019) for developing a high-resolution method termed **LipidArrays** that will enable the diversity and number of lipids in biological samples to be characterized. They are particularly focusing on cell development, cardiovascular disease and dementia in this study.

Precision medicine

- In neuroscience and mental health, Cardiff University researchers are integrating genetic data with 'omics data (expression, protein, epigenetic, regulatory) in an efficient manner to uncover the biological pathways underlying phenotypically complex disorders. They are also linking detailed phenotypic data to genomics data in order to identify novel, genetically homogeneous patient strata (cutting across diagnostic criteria), and to discover how the relevant biological pathways interact with environmental factors (diet, smoking etc.) in these strata to influence health, thereby leading to drugs and/or interventions.
- Large projects at planning stage include analysis of samples from the paediatric neonatal MRC **RHINO cohort** and **HealthWise Wales**, to address the role of metabolites and lipids in causing or contributing to disease processes, how to use their detailed profile, e.g. measuring thousands simultaneously in plasma or tissue (which will change dynamically unlike our genome) for diagnosis and patient stratification, for monitoring treatment success and outcome, and how much of our metabolome/lipidome is heritable vs environmentally controlled.

Population and Public health

- The Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement (**DECIPHer**) is a UKCRC Public Health Research Centre of Excellence. It is a Cardiff University-led (Professor Simon Murphy) strategic partnership with the University of Bristol (Professor Rona Campbell) and Swansea University (Professor Ronan Lyons). DECIPHer includes research programmes on data linkage and natural experiments which are led by Professor Ronan Lyons at Swansea University, and works closely with the Farr@CIPHER.
- The **Centre for Trials Research** (also mentioned in a previous section) designs and delivers multi centre studies across the UK and beyond.

Disease areas

- Cardiff University's research mainly focuses on: mental health; neuroscience; infection, immunity and inflammation; cancer; and population health.

Analytics

- Cardiff University researchers have developed methods to:
 - integrate genetic data with multiple 'omics data across neurodegenerative and neuropsychiatric disorders;
 - analyse bacterial genomic data, risk models for population health risk factors, and informatics approaches to enable data and software sharing, standardisation and reuse through MRC CLIMB; and
 - process massive lipidomic/metabolomics datasets.

Collectively these activities are focused on delivering tools that can be accredited and adopted into NHS service.

- Within the **Division of Population Medicine**, biostatistics methodology is focussed on the highly-interconnected areas of longitudinal data analysis, missing data and causal inference for a fuller, deeper and more insightful assessment of the assumptions underlying the most commonly used biostatistical techniques, as applied to the epidemiological analysis of large-scale record-linked health and education datasets.

e-Infrastructure

Compute

- Cardiff University has a computational infrastructure that comprises over 16,000 processing cores, more than 77TB of total RAM and over 3PB of total data storage. They are currently in the process of procuring >£3M of hardware to update and renew their HPC systems, to carry them through into the 2020s.
- Cardiff University researchers have access to the **Supercomputing Wales** infrastructure, which provides around 17,000 cores across eight sites based on a hub and spoke model with hubs in Swansea and Cardiff.
- The **Cloud Infrastructure for Microbial Bioinformatics** (CLIMB) includes a distributed four-site computing system capable of supporting over 1,000 virtual bioinformatics servers. This system consists of hardware with 7,680 cores, 2,304 TB of cross-site replicated storage and 78TB of RAM.
- Through its **GW4 Alliance** membership, CU will have access to a new (to be installed in 2017) Tier 2 HPC service for UK-based scientists. The system will be one of the world's first to be based on Broadcom's Vulcan server-class chip.

Training

- Opportunities for skill development are available through Cardiff University's existing and planned training programmes including: undergraduate and postgraduate teaching modules covering bioinformatics, statistics and programming; a research methods consultancy service for the NHS and social care in SE Wales; training in the use of the supercomputing facility; and **MSc in Bioinformatics**; and **MSc in Bioinformatics and Genetic Epidemiology**. Furthermore, curriculums have been designed to enable the teaching of module elements to meet continuing professional development and training needs.

Public Involvement/Engagement

- The Centre for Trials research is experienced in undertaking public involvement and engagement activities relating to the retention and use of data.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@CIPHER (led and based in Swansea University) – award partners include the Universities of: Cardiff, Bristol, Brighton, Oxford, Exeter and Surrey, Western Australia, Monash, Ottawa; and the Welsh Government; and the Finnish National Public Health.
- Supercomputing Wales consortium (hubs at Cardiff and Swansea Universities)
- GW4 Alliance University partners: Bath, Bristol, Cardiff and Exeter
- MRC Medical Bioinformatics CLIMB – encompasses the following partners – University of Warwick, Cardiff University, Swansea University, University of Birmingham, University of Bath, University of Leicester and The Quadram Institute (a BBSRC research institute)
- UK Biobank collaboration

NHS and other government organisations

- Genomics England – leading on the Wales Genomic Medicine Centre
- Public Health Wales; Cardiff and Vale NHS trust and other NHS Trusts across Wales.

Swansea University

Top-level insights

- Swansea University has an established international reputation as a leading centre in the field of health informatics, built upon substantial developments over a long period of time and a large network of collaborations. It is this track record that Swansea University's Sol is based on; including, the activities of the Farr@CIPHER (Swansea) and those of other aligned investments in Wales including: the Secure Anonymised Information Linkage (SAIL) Databank; the UK Secure e-Research Platform (UKSeRP) and its sister technology National Research Data Appliance; ADRC-Wales and the MQ Adolescent Mental Health/ MRC Mental Health Data Pathfinder initiatives.
- Its main strengths are in: 1. combining health and administrative data at scale and from a variety of sources to increase understanding of population health and social well-being that will have an impact on policy and practice; 2. creating innovative solutions to problems in accessing data, privacy protection, linkage, analysis, interoperability and researcher access; 3. developing data sharing and analysis platforms.
- Swansea University's considerable expertise in handling and using routine data comes through 10 years' experience in the development of the SAIL Databank and its expansion to support multiple national initiatives, mainly through the UKSeRP.
- UKSeRP is a modularised platform able to provide anything from a desktop environment to full supercomputing facilities with a wide range of analytical tools on board. Hence, it has brought scalability and reduced the complexity and risks of providing IT infrastructure and analytical support for a number of national research initiatives, achieved through using spare capacity of established proven systems.
- Swansea University has strong links and data sharing agreements for access to anonymised person-based health and social care data for research with multiple ministries in the Welsh Government; all ten health boards and trusts; 80% of general practices; plus a growing number of third sector organisations. The SAIL Databank provides secure access to billions of this data about the population of Wales. This has positioned Wales as amongst the deepest phenotyped nations, supporting some of the largest scale UK observational and interventional research in this field.
- Swansea University training courses in informatics include: the first MSc in Health Data Science; and the sole internationally respected Introduction/ Advanced Analysis of Linked Health Data courses in the UK.

Strengths, Capabilities and Opportunities

Strengths

Informatics Investments

- Swansea University has a long track record of investments in health informatics that include the development of: the **Welsh Health Information Research Unit** in 2006; and the subsequent creation of the Secure Anonymised Information Linkage (**SAIL**) **Databank**, which in turn underpinned the successful awards in the **Farr@CIPHER** (*Director: Professor Ronan Lyons*); and **ADRC-Wales** (*Director: Professor David Ford*). These investments, together with closely associated research groups, the NHS and industry, are housed at the **Data Science Building** – a six floor 3,000 m² purposely designed secure data facility, funded by an £8M investment from MRC, ESRC and the Welsh Government – which opened in 2016 with a lifespan of at least 25 years.

- The **SAIL Databank** is managed by the Health Informatics Group, which is part of the Patient and Population Health Informatics research theme of the Institute of Life Sciences, Swansea University Medical School. SAIL is a valuable and unique National Data Repository which already supports a wide variety of e-health research and evaluation projects, and which is increasingly becoming recognised as a valuable resource for healthcare policy making, service planning and evaluation (*see also data section*). Backed and endorsed by the Welsh Government, the **SAIL Databank** receives core funding from the Welsh Government's Health and Care Research Wales. The current SAIL has secured renewal funding to 2020.
- As part of the **MRC Medical Bioinformatics investment into the Cloud Infrastructure for Microbial Bioinformatics** (CLIMB; £8.5M), Swansea is one of four distributed sites (Warwick, Birmingham, Swansea and Cardiff). It provides infrastructure based on OpenStack, with a variety of virtual server infrastructure provided to the research community. In addition to providing traditional data storage and compute infrastructure, this provides a large geo-replicated object store both to support longer-term storage and to underpin collaboration. The project is working on the automated provisioning of genomic software and pipelines to assist the microbial genomics community. CLIMB currently supports over 1,000 microbiological researchers across the UK (*see also section 2.1.2*)
- Swansea University has a strategy of continuing investments in internationally-leading interdisciplinary research that include: the new **Computational Foundry** (£31M – opening in 2018); the **Academy for Advanced Computing**; and investments in medical engineering through the **Zienkiewicz Centre for Computational Engineering** and the joint **Engineering/Medicine Centre for NanoHealth**.

Knowledge/Expertise

- There is a variety of backgrounds and expertise, including: epidemiology, statistics, health service research, public health science, clinicians (multiple specialties), geography, demography, social science, economics, law, information governance, public engagement, genetics, bioscience, computer science, mathematics, and physics.
- With ten years' experience operating the SAIL Databank, SU researchers have developed an extensive knowledge on how to build, implement and operate industrially-scaled socio- technical data platforms.
- The Swansea University-led (*Jones*) Information Governance projects are:
 - the **GeDI** project – on incorporating genetic data into data safe havens;
 - the **MORPHeD** project – on the use of integrated smartphone and health data; and
 - the **ACoRN** project – on scoping barriers and drivers to cross-centre research.

Data

- Backed and endorsed by the Welsh Government, the **SAIL Databank** receives core funding from the Welsh Government's Health and Care Research Wales. Currently, there are 50 (if include the GP practices) routine and 115 project-specific data flows involving all data modalities, with over 19B recordings on 5.1M living and dead individuals, making Wales among the most deeply phenotyped regions of the UK.

SAIL Databank is increasing its holdings in health determinants and outcomes data with sharing agreements in place with organisations such as South Wales Police, the National Offenders Management Service, local government social services and housing departments and many more.

- The **Wales Electronic Cohort for Children (WECC)** (*Lyons* lead investigator) is the first and largest (>1,000,000) complete child cohort in the UK, aimed at investigating the widest possible range of social and environmental determinants of child health and social outcomes by exploiting the potential of routinely collected datasets.

Compute

- The **Academy for Advanced Computing** supports researchers across the University in exploiting HPC technology available through the **HPC Wales** supercomputing infrastructure.

Translational and innovation capabilities

Links with NHS

- The **Prudent Healthcare Intelligence Hub**, based at SU but working across the whole of the Welsh NHS, delivers analytical services to support the direct needs of healthcare organisations and policymakers by undertaking high impact analyses and evaluations. With involvement of staff from across the Welsh NHS, this cooperative model, underpinned by Farr Institute infrastructure and SAIL Databank data, is proving a highly efficient means of gaining real impacts into society, as well as building capacity and capability in the NHS. In the first year of operation, nine projects have been developed with Abertawe Bro Morgannwg, Aneurin Bevan, Hywel Dda and Cwm Taf University Health Boards, the Welsh Ambulance Service and Public Health Wales NHS Trusts. These projects cover a wide range of activities and diseases areas including physical activity, injuries, cancer, and cardiovascular disorders.

Opportunities

- There is enthusiasm for engagement in **citizen-driven health research** from each of the applied clinical research groups. For example, the Asthma UK CAR team is particularly interested in capturing data from smart inhalers, wearables (including biological sensors) and mobile apps. The UK MS Register web portal collects PROMs data from 15,000 participants and plans to add complete actigraph data dumps from mobile phones to assess utility.

Swansea/Wales is an attractive test-bed site, as wearable technology develops, given the available operational data sharing platforms. Furthermore, the Chancellor supported the **Swansea Bay City Regional Deal** (£1.3B) in the 2016 Budget. This envisages the development of **'The Internet Coast'** with a new transatlantic fibre-optic cable from New York to Oxwich (Swansea) and next generation wireless (5G) supporting the 'Internet of Energy (Tidal Lagoon) and **'The Internet of Health and Wellbeing'**, providing a major boost for data intensive research opportunities.

- Swansea University is a core member of the **ARCH** partnership (**a Regional Collaboration for Health**) to enable an enhanced multi-sectoral data linkage on a population of 1M in South West Wales being used to drive the case for change in modernising health services, including the development and evaluation of a Wellbeing Programme.

These unique ambitions will be enhanced by a coherent, investor-attractive digital economy developed through the Swansea Bay City Deal programme. Building on these developments, SU has plans for collaboration with Manchester and other universities on the Connected Health Cities and regions agenda.

Potential gaps

- Swansea University has modest expertise in advanced health data analytics; and has started to form links with health data/technology industry.

Other relevant capabilities

Research themes

Swansea University has niche, but growing, focused areas of expertise in discovery science and precision medicine with Swansea University researchers involved in several multi-national multi-omics data linkage studies.

Discovery Science

- Swansea University is involved in active collaborations focussed on **epi-genome association studies** including:
 - asthma and allergy with Imperial
 - epilepsy and brain disorders with the NIH Epi4K Exome Consortium
 - NIH Epi25 Consortium with the Broad Institute
 - Wellcome Trust Synapopathy Consortium with UCL
 - BioJume Consortium with KCL
 - the EpiGen Global Consortia with Oxford, Dublin and UCL.

Precision medicine

- In the **Centre for NanoHealth** investigators are engaged in research involving semi-conductor processing, biosensor technologies, microfluidics and microneedles. This includes the development of generic epitaxial graphene biosensors for ultrasensitive detection of cancer risk biomarkers, point of care testing for oxidative stress, cardiac markers and hormones using packaged silicon devices and electrochemical chip designs.
- Researchers from the **Zienkiewicz Centre for Computational Engineering** are involved in a broad range of medical research focussing on cardiovascular and respiratory flow modelling and the development of pulse wave screening devices for aortic aneurysms in collaboration with a major automotive engineering company.

Population and Public health

- Patient & Population Health and Informatics is one of four main research themes in Swansea University Medical School. Furthermore, it links Farr@CIPHER, ADRC-Wales and **National Centre for Population Health and Wellbeing Research** (NCPHWR) structures and personnel and embeds methodological developments in accessing, interpreting and analysing data around focussed health questions.
- Research spans the entire range of study designs from the creation of population based cohorts (focussed on demographic subgroups, diseases and injuries, and settings, such as children, housing, schools, or criminal justice), hybrid cohorts (embedded traditional cohorts within larger population e-cohorts), exploiting the UK's rich cohorts, measurement of population burden of disorders, evaluations of natural experiments and routine data enhanced individual/cluster randomised controlled trials.
- Swansea University has developed a clinical trials module (**RITREVA**) which swiftly and reliably produces ranked lists of patients eligible for trials, massively speeding up recruitment processes. Welsh Government is currently considering a proposal to rollout this proven prototype across Wales to produce a national trial recruitment platform.

Disease areas

- Swansea University has research programmes on a number of disease areas with particular focus on: injury, environmental exposures, substance use, mental health, infection and child health, as well as the traditional major chronic diseases.

Analytics

- There are active collaborations with several consortia on remote access to multi-modal data and development of new analytical methods, e.g. time-varying confounding in longitudinally linked data (UCL), multi- omics/phenome analyses (Imperial, Cambridge) and image analysis (Oxford, UCL).
- Leading edge probabilistic and deterministic **linkage** allows multiple data sources to be linked together, and the process to be tailored based on the quality of achievable linkage.
- Swansea University has growing capability in **natural language processing** (NLP). As well as supporting and making available leading open-source tools, Swansea University has invested in three main products:
 - **IBM Content Analytics / IBM WATSON** – a large scale data centre infrastructural deployment bringing in IBM latest content management systems to feed the WATSON Content Analytics system. This system provides an end-to-end ingestion and analysis platform for performing analysis on unstructured information.
 - Free-text to **SNOWMED-CT3 engine** – embedded in a number of NRDA systems this engine will allow for the conversion of text, office documents and PDFs into HL7 messages to ingest, convert and annotate source documents with SNOWMED-CT3 terms. The system also then has the ability to query this dataset and, using a rules-based engine, derive new variables which can form a traditional dataset for onward transmission.
 - **GATE** has been installed in version 2 of the National Research Data Appliances to create a pipeline for derived deep phenotypic data from NHS text based source documents to be added to SAIL
- Oxford University has located the DPUK central processing hub for **image analysis** with Swansea University to provide common shared infrastructure connecting multiple imaging centres and supporting cross-centre collaboration. This is based on geo-replicated OpenStack computer clusters and storage subsystems.
- Swansea University also offers access to centralised systems and a wide variety of **software** stacks to support the needs of a variety of national programmes and projects, e.g: Transmart, LabKey, MACRO 4, RedCap, XNat, JIRA.

e-Infrastructure

Compute

- Swansea University researchers have access to the **HPC Wales supercomputing** infrastructure, which provides around 17,000 cores across eight sites based on a hub and spoke model with hubs in Swansea and Cardiff.
- The **Cloud Infrastructure for Microbial Bioinformatics** (CLIMB) includes a distributed four-site computing system capable of supporting over 1,000 virtual bioinformatics servers. This system consists of hardware with 7,680 cores, 2,304 TB of cross-site replicated storage and 78TB of RAM. Through this infrastructure, Swansea University has access to: 2,000 cores, 500TB storage, 40TB memory and 1PB object storage.
- Swansea has invested in a range of systems underpinning UKSeRP and SAIL offering unique capabilities selected to best fit the needs of individual projects and users, including IBM DB2, Microsoft SQL Server, Postgres, HADOOP, Elastic Search, and OpenStack/Linux with IBM Watson/Content Analytics, CliniThink and GATE supporting NLP. Across all warehouses, databases, VM/VDi, imaging and genomics infrastructure components researchers have access to 2,328 cores, 35 TB of memory and 3.4 PB of storage.

Platforms

- MRC Farr Institute capital funding enabled Swansea University to decouple platform components, making them available as “white label” technologies to support others in their pursuit of establishing robust and acceptable data sharing environments in different jurisdictions or across multiple research partners. Notable amongst these are Swansea University’s **National Research Data Appliances (NRDA)**, a sophisticated, standalone “SAIL Databank in a box” scalable technology set, designed to manage multiple datasets within a complex data holding organisation (e.g. NHS Trust or a Local Authority) and create research ready lined datasets. By automating the production of data quality reports and standardised metadata, probabilistic demographic linkage, the NRDA make simple the task of managing complex data for data owning organisations, benefiting them and those that have agreed data sharing arrangements in place. NRDA are currently in use in Wales and are being considered by several other parts of the world. Modular in design, the NRDA can be supplied with on-board national language processing facilities, allowing data owners to continuously “read in” their free text documentation (such as letters and summaries) and create SNOMED CT codes for analysis and sharing.

Trusted Research Environments

- MRC investment in UKSeRP extended NRDA functionality to include federation of NRDA and the operation of a cloud-like ISO2001 accredited data sharing platform service to research programmes, with tailored governance models and devolved management controls removing the need for capital investment and development time. Requirements of ARDC-Wales have led to an instance of UKSeRP with enhanced security controls to meet the expectations of UK Government departments. UKSeRP is now used by 40+ cohorts in DPUK, UK Biobank, ALSPAC, East London Genes for Health and the UK MS Register. The technology is being shared with a number of international partners to support global research initiatives.
- Swansea University is also part of the JISC SafeShare initiative to provide higher assurance connectivity between institutions and safe settings and has both Farr and ADRC connections to relevant secure networks to facilitate secure cross-centre access and collaboration.

Training

- Swansea University is a long-term provider of high quality education, training and CPD activities in the field and is the sole UK provider of the internationally respected **Introduction/Advanced Analysis of Linked Health Data courses**, originated by Prof D’Arcy Holman of UWA, now delivered as SU/UWA collaboration.
- Courses are available as a standalone basis, or part of an **MSc in Health Informatics**. NHS and industry internships form part of the dissertation and work-based project of this course. In partnership with the NHS Wales Informatics Services and Health Boards in Wales, this course is able to offer NHS research opportunities within the NHS Wales Informatics Research Laboratories based at Swansea University. Over 100 students have now registered for the MSc/MRes and 238 professionals have completed individual modules as part of their continuous professional development.
- Swansea University also offers UK’s first **MSc in Health Data Science**.
- An extensive programme of other research- relevant short courses, including **Open University** modules in Health Informatics, is provided each year to students from across the world.
- Swansea University makes major contributions to the **Farr Institute**’s PhD and ECR symposia.

Public Involvement/Engagement

- The Farr@CIPHER, ADRC-Wales and SAIL have a joint active Public Engagement programme, involving consumers fully in all relevant stages of their work by:
 - facilitating the operation of a strong and inclusive Consumer Panel for Data Linkage Research, to advise, support and advocate their work;
 - involving patients, service users, carers and the general public in their research;
 - building and maintaining a web-based public information resource that showcases the benefits of data linkage research;
 - hosting and participating in, regular events designed to be of interest to members of the public.

Partnerships

Key strategic interconnections with health data research and innovation related investments at a glance:

Research Councils and other ROs

- Farr@CIPHER (led and based in Swansea University) – award partners include the Universities of: Cardiff, Bristol, Brighton, Oxford, Exeter and Surrey, Western Australia, Monash, Ottawa; and the Welsh Government; and the Finnish National Public Health. Through this investment, strong links to the other Farr Centres: Farr@London (UCL), Farr@HeRC (Manchester) and Farr@Scotland.
- MRC Medical Bioinformatics CLIMB – award partners include: University of Warwick (lead); Swansea University; University of Birmingham; and Cardiff University
- Supercomputing Wales consortium (hubs at Cardiff and Swansea Universities)
- Strong links with ARDC-Wales and through this to the other ADRCs (England, Northern Ireland and Scotland) and the ADRN
- MRC-led (Cardiff/Bristol/Swansea) DECIPHer
- MRC Dementias Research Platform UK
- Imperial's MRC Medical Bioinformatics award collaboration
- UK Biobank's Outcome Adjudication team (Oxford, Edinburgh)

NHS and other government organisations

- Multiple ministries in Welsh Government
- All ten health boards and trusts
- Welsh Government's Health and Care Research Wales Board

Data providers

- 80% coverage of GPs in Wales
- Public Health England Wales

Key strategic interconnections with health data research and innovation related investments at a glance *(continued)*:**Industry**

- Technology companies: IBM, Fujitsu

International

- International Population Data Linkage Network
- ICES in Ontario Canada
- Monash University
- Norwegian National Research Council
- NIH Epi4K Exome Consortium
- NIH Epi25 Consortium



ANNEXES

ANNEX 1

Biomedical Research Centres research themes

(as presented on the NIHR website)

BRCs host research themes across a range of disease and therapeutic areas and each BRC has substantial portfolio either across a range of clinical and research areas, or in a specific clinical or research area. Details of the BRCs and their research themes are presented in the table below.

BRC Name	Name including host organisation and academic partner	Research themes	Funding for 5 years from 1 April 2017
NIHR Barts Biomedical Research Centre	NIHR Biomedical Research Centre at Barts Health NHS Trust and Queen Mary University of London	Cardiovascular devices and innovative trials, inherited cardiovascular disorders.	£6,557,380
NIHR Birmingham Biomedical Research Centre	NIHR Biomedical Research Centre at University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham	Inflammatory arthritis, inflammatory bowel disease, inflammatory sarcopaenia.	£12,120,962
NIHR Bristol Biomedical Research Centre	NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol	Cardiovascular disease, mental health, nutrition, diet and lifestyle (including obesity), reproductive and perinatal health, surgical innovation.	£20,858,545
NIHR Cambridge Biomedical Research Centre	NIHR Biomedical Research Centre at Cambridge University Hospitals NHS Foundation Trust and the University of Cambridge	Antimicrobial resistance, cancer, cardiovascular and respiratory disease, dementia and neurodegenerative disease, gastrointestinal disease, integrative genomics, mental health, metabolism, endocrinology and bone, neuroscience, nutrition, diet and lifestyle, population and quantitative science, transplantation and regenerative science, women's health and paediatrics.	£114,300,000
NIHR Manchester Biomedical Research Centre	NIHR Biomedical Research Centre at Central Manchester University Hospitals NHS Foundation Trust and the University of Manchester	Advanced radiotherapy, cancer prevention and early detection, cancer precision medicine, dermatology, hearing health, respiratory medicine, targeted therapy in musculoskeletal diseases	£28,500,000

BRC Name	Name including host organisation and academic partner	Research themes	Funding for 5 years from 1 April 2017
NIHR Great Ormond Street Hospital Biomedical Research Centre	NIHR Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London	Advanced treatments for structural malformation and tissue damage, gene, stem and cellular therapies, genomics and systems medicine, novel therapeutics and their translation into childhood disease.	£37,005,790
NIHR Guy's and St Thomas' Biomedical Research Centre	NIHR Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London	Cardiovascular disease, cutaneous medicine, genomic medicine, imaging sciences, infection and immunity, oral health, regenerative medicine and cellular therapy, transplantation, women and children's health.	£64,400,267
NIHR Imperial Biomedical Research Centre	NIHR Biomedical Research Centre at Imperial College Healthcare NHS Trust and Imperial College London	Brain sciences, cancer, cardiovascular, gut health, immunology, infection and AMR, metabolic medicine and endocrine, surgery and technology	£90,008,746
NIHR Leeds Biomedical Research Centre	NIHR Biomedical Research Centre at Leeds Teaching Hospitals NHS Trust and the University of Leeds	Preventing disease and disability in immune mediated inflammatory disease, improving treatment of osteoarthritis	£6,736,575
NIHR Leicester Biomedical Research Centre	NIHR Biomedical Research Centre at University Hospitals of Leicester NHS Trust and the University of Leicester	Cardiovascular, lifestyle, respiratory.	£11,591,314
NIHR Maudsley Biomedical Research Centre	NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London	Affective disorders and interface with medicine, bioinformatics and statistics, biomarkers and genomics, child and neurodevelopmental disorders, clinical and population informatics, dementia and related disorders, mobile health, neuroimaging, obesity, pain, patient and carer involvement and engagement, psychosis and neuropsychiatry, substance use, translational therapeutics.	£65,977,500

BRC Name	Name including host organisation and academic partner	Research themes	Funding for 5 years from 1 April 2017
NIHR Moorfields Biomedical Research Centre	NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology	Gene therapy, genomic medicine and informatics, inflammation and immunotherapy, regenerative medicine and pharmaceuticals, visual assessment and imaging.	£19,075,000
NIHR Newcastle Biomedical Research Centre	NIHR Biomedical Research Centre at Newcastle upon Tyne Hospitals NHS Foundation Trust and Newcastle University	Dementia, liver disease, musculoskeletal disease, neuromuscular disease, skin and oral disease.	£16,208,633
NIHR Nottingham Biomedical Research Centre	NIHR Biomedical Research Centre at Nottingham University Hospitals NHS Trust and the University of Nottingham	Deafness and hearing problems, gastrointestinal and liver disorders, mental health and technology, musculoskeletal disease, respiratory disease.	£23,642,003
NIHR Oxford Biomedical Research Centre	NIHR Biomedical Research Centre at Oxford University Hospitals NHS Foundation Trust and the University of Oxford	Antimicrobial resistance and microbiology, cardiovascular, diabetes and metabolism, gastroenterology and mucosal immunity, genomic medicine, haematology and stem cells, multi-modal cancer therapies, multimorbidity and long term conditions, musculoskeletal, neurological conditions, obesity, diet and lifestyle, respiratory, stroke and vascular dementia, surgical innovation and evaluation, technology and digital health, vaccines for emerging and endemic diseases.	£113,718,800
NIHR Oxford Health Biomedical Research Centre	NIHR Biomedical Research Centre at Oxford Health NHS Foundation Trust and the University of Oxford	Adult mental health, older adults and dementia, precision psychological treatments	£12,824,900
NIHR Royal Marsden Biomedical Research Centre	NIHR Biomedical Research Centre at The Royal Marsden NHS Foundation Trust and The Institute of Cancer Research	Breast cancer, gastrointestinal cancers, novel cancer therapeutics, prostate cancer, targeted physical therapies, uncommon cancers.	£43,074,315

BRC Name	Name including host organisation and academic partner	Research themes	Funding for 5 years from 1 April 2017
NIHR Sheffield Biomedical Research Centre	NIHR Biomedical Research Centre at Sheffield Teaching Hospitals NHS Foundation Trust and the University of Sheffield	Translational neuroscience for chronic neurological disorders.	£4,049,681
NIHR Southampton Biomedical Research Centre	NIHR Biomedical Research Centre at University Hospitals Southampton NHS Foundation Trust and the University of Southampton	Life-course nutrition, lifestyle and health, respiratory and critical care.	£14,509,067
NIHR University College London Hospitals Biomedical Research Centre	NIHR Biomedical Research Centre at University College London Hospitals NHS Foundation Trust and University College London	Cancer, cardiovascular disease, deafness and hearing, dementia and mental health, immunity, inflammation and immunotherapeutics, neurological diseases, obesity, oral health and disease.	£111,503,317

ANNEX 2

Further information on ESRC investments in Big Data mentioned in section 2.5 of this report:

1. Business and Local Government Data Research Centres

Since January 2014, ESRC have been investing in Business and Local Government Data Research Centres. These Data Research Centres make data, routinely collected by business and local government organisations, accessible for academics in order to undertake research in the social sciences of mutual benefit to data owners and in ways that safeguard individuals' identities. The Centres, along with the ADRN, will make a significant contribution towards ensuring the future sustainability of the UK research competitiveness, supporting the UK in maximising its innovation potential and driving economic growth. The Centres are funded until December 2019. The Directors of these Centres are:

- Urban Big Data Centre: Professor Piyushimita (Vonu) Thakuria at the University of Glasgow
- ESRC Business and Local Government Data Research Centre: Professor Vania Sena at University of Essex
- **Consumer Data Research Centre:** Professor Mark Birkin at University of Leeds and Professor Paul Longley at University College London

2. RCUK Digital Economy Hubs

To maintain momentum in the new and emerging forms of data area the ESRC awarded £1.4M of funding to the RCUK Digital Economy Hubs in 2014. This funding was to build on and/or enhance work they have been carrying out which is closely aligned to this agenda, as well as to conduct complementary activities which will contribute to the development of infrastructure for research using social media data. This work was funded for a period of 18 months running from February 2014 to August 2015.

3. Civil Society Data partnership Projects

In 2014 the ESRC commissioned four Civil Society Data Partnership Projects, aimed at establishing or building on relationships between academic researchers and civil society organisations to demonstrate the value of improved data infrastructure, enabling collection and analysis of data which is of interest to civil society organisations and through enabling the sector to better utilise its own data. 4 projects were funded:

- Utilising Big Data in the Practice of Torture Survivors' Rehabilitation.
- A Profiler for Crime, Criminal Justice and Social Harm
- Data Resource Construction: Open Data, Grantmaking Data, and the Organisational and Financial Base of the Third Sector
- Developing the Use of Administrative Data on Scotland's Civil Society

4. Centre for Doctoral Training in New Forms of Data

In 2016 ESRC commissioned a Centre for Doctoral Training (CDT) in new forms of data to be developed in collaboration with non-academic partners and will focus on real-time data analysis, data linkage and interoperability of data from different sources. The substantive focus is on 'new forms of data' in a broad sense and on the opportunities provided by these data to study social processes as they unfold.

This CDT was awarded to the Data Analytics and Society CDT led by the University of Leeds in partnership with the Universities of Manchester, Sheffield and Liverpool.

The Centre will focus on

- Promoting the creation and analysis of new longitudinal and streamed data resources for socio-economic investigations.
- Creating new methods (e.g. scaling up existing methods for real time big data analytics),
- Investigating social processes (e.g. virtualisation of retailing; data-driven decision making and social behaviours)
- Facilitating interventions (e.g. resource targeting, network planning, social media apps for diet, travel, lifestyle planning).

The CDT will support minimum of 10 students per year for three years with the first cohort starting in October 2017.

5. New and Emerging Forms of Data Policy Demonstrator Projects

In January 2017 ESRC commissioned policy demonstrator projects to showcase the potential value of new and emerging forms of data for policymakers. 8 projects were funded as part of this call running from February 2017 for up to a year. The projects funded in the call are listed below.

- Social sensing of health and wellbeing impacts from pollen and air pollution
- HABITS: improved policy to mitigate pollutant and inactivity related health burdens through new big data
- Data Awareness for Sending Help (DASH)
- The food sentiment observatory: exploiting new forms of data to help inform policy on food safety and food crime risks
- Population247NRT: near real-time spatiotemporal population estimates for health, emergency response and national security
- Inclusive and healthy mobility: understanding trends in concessionary travel in the West Midlands
- New and emerging forms of violence data for crisis response: a comparative analysis in Kenya
- Centre for cyberhate research and policy: real-time scalable methods and infrastructure for modelling the spread of cyberhate on social media

ANNEX 3

List of abbreviations

ADRC	Administrative Data Research Centre
ADRN	Administrative Data Research Network
AHSC	Academic Health Science Centre
AHSN	Academic Health Science Network
AMR	Antimicrobial resistance
ATI	Alan Turing Institute
BBSRC	Biotechnology and Biological Sciences Research Council
BRC	Biomedical Research Centre
CCG	Clinical Commissioning Group
CPRD	Clinical Practice Research Datalink
CPU	Central Processing Unit
CR UK	Cancer Research UK
DHSC	Department of Health and Social Care
DPUK	Dementia Platform UK
DRI	Dementia Research Institute
EMBL	European Molecular Biology Laboratory
EHR	Electronic Health Record
EPR	Electronic Patient Record
EPSRC	Engineering and Physical Sciences Research Council
ESRC	Economic and Social Research Council
EU	European Union
GB	Gigabyte
GeL	Genomic England Ltd
HIC	Health Informatics Collaborative
HPC	High Performance Computing
ICT	Information and Communications Technology
IG	Information Governance
IMI	Innovative Medicines Initiative
MRC	Medical Research Council
NHS	National Health System
NICE	The National Institute for Health and Care Excellence (UK)
NIH	National Institutes of Health (USA)
NIHR	National Institute for Health Research (UK)
ONS	Office for National Statistics (UK)
OS	Operating System
PB	Petabyte
PHE	Public Health England
PI	Principal investigator
TB	Terabyte

Medical Research Council (Swindon office)

2nd Floor David Phillips Building
Polaris House
North Star Avenue
Swindon
SN2 1FL

Medical Research Council (London office)

14th Floor
One Kemble Street
London
WC2B 4AN

Phone (+44) (0)1793 416200

[mrc.ukri.org](https://www.mrc.ukri.org)

The Medical Research Council is part of UK Research and Innovation