Medical Research Council

The Opportunity of Biomedical Data Science

Maximising the opportunities from data science for innovative biomedical research

MRC strategic review

Scope and purpose of this review

The UK has an extensive landscape in both biomedical and data science, with a number of areas of intersecting and complementary activity[1]. However, the recent MRC Unit and Centre Portfolio Review highlighted data as an essential and shared priority across all identified priority health needs, and noted a number of key areas of opportunity and need[2].

Data science[3] is an integral part of biomedical research and will be referred to in this report as 'biomedical data science'. Given technology and other advances in biomedical research, there is an enormous and yet to be harnessed potential of biomedical data science, with a risk that the UK gets left behind in the global advance of data science in innovative biomedical research[4]. This review was initiated to seek recommendations to address this challenge.

An expert group (see Annex 1) was convened and invited to identify opportunities and needs in biomedical data science and provide key recommendations to the MRC. The review sought to consider the full spectrum of quantitative biomedical data science interests, spanning all research disciplines and the full range of approaches, ranging across mathematics, statistics and computing. In addition to their own expertise and knowledge of the landscape, members drew on evidence from the MRC Unit and Centre Portfolio Review, information on major relevant activities and investments and written submissions from key organisations[5], and were cognizant of other reviews and strategies in development within MRC and across UKRI on digital research infrastructure and training and skills[6].

This report sets out the identified opportunities, enablers and required interventions for effective data science in biomedical research.

[1] Major UK investments in data science, biostatistics and health data research include those named in footnote 5; as well as a number of interdisciplinary data science university institutes and data resources such as UK Biobank, Genomics England, longitudinal population cohorts.
[2] The optimal use of data requires infrastructure, technology and expertise in gathering, curation, sharing, integration and analysis, across multiple sources. There is a need for developments in bioinformatics and modelling, including temporal and causal inference methods, and visualisation. Further themes, see Annex 2.
[3] For the purposes of this review, broad working definitions were used. **Data science**: the use of scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, including data analysis, data integration and interfaces with other data domains (e.g. health and population data), systems/computational modelling, new analytical methods, adaptation of AI methods and AI-enabled analytics, data wrangling, research software engineers, data engineers, data itself and digital infrastructure. **Biomedical data**: to include data ranging from omics and microscopy to medical imaging and population cohort data.
[4] A leading example of a high profile, successful activity in this field is the US Broad Institute, which pioneers large data approaches in biology, as evidenced by their recent investment in a new Centre for machine learning in biology https://www.broadinstitute.org/news/broad-institute-launches-eric-and-wendy-schmidt-center-connect-biology-machine-learning.
[5] Selected relevant UK organisations were invited to outline their views on major needs, opportunities and challenges in data science. Responses were received from the Alan Turing Institute, Dementia Research Institute, Francis Crick Institute, Health Data Research UK, MRC Biostatistics Unit, MRC Human Genetics Unit, MRC Laboratory of Molecular Biology, MRC London Institute of Medical Science, MRC Precision Medicine Group, Wellcome Sanger Institute
[6] A number of documents are under development, including the UKRI Digital Research Infrastructure Strategy, with the recent UKRI infrastructure roadmap a published document of relevance.

1 Biomedical data science is a substantial and urgent opportunity

Biomedical data science has the capacity to answer fundamentally important questions in human health and disease. The ability to integrate and interrogate data, across scales, modalities and systems offers the chance to enhance understanding of disease aetiology and develop new interventions that improve health. Notwithstanding existing investments[7], the scale, speed and breadth of these emerging opportunities have resulted in gaps emerging in UK capability to maximise the opportunities from data science for innovative biomedical research. The potential growth of this area over the next 5-10 years is already recognised by the industry sector, with the opportunity to leverage substantial investment from the pharmaceutical industry, increasing interest from "big tech" and emerging biotech companies.

1.1 Opportunity for impact by integrating multiple diverse data types and sources

The biggest opportunities for impact from biomedical data science will come through consideration of all relevant datasets in a '360 degree' approach, including all data types with relevance to human health in the broadest sense.

The impact data science has had on COVID-19 has demonstrated that relevant data spans all scales and types, including **molecular, population, behavioral, social, environmental and geographical data;** and spans spectrums of human, animal, societal and environmental health (Figure 1).

With appropriate safeguards, data may be collected by a variety of means including analytical instruments, imaging (e.g. microscopes, and scanners), 'omic technologies, electronic health records, questionnaires, smartphones and wearables, diagnostic and medical devices, satellites, sensors; from daily use, from routine clinical practice or in observational or interventional studies and trials.

Improved longitudinal data analysis approaches will unlock the value of these data types, including within the UK's well curated large-scale cohorts and biobanks. Increasingly there is the potential to **use non-health data,** such as financial, transactional, demographics, social deprivation, lifestyle and risk factor information **to understand health impact.** There are particular opportunities and challenges in the integration and use of data that are multidimensional, multimodal or multiscale.

1.2 Exemplar cases

Many examples of the potential for biomedical data science exist, and the following are illustrative or exemplar cases where there is potential for high and timely impact:

- **Imaging**. Delivering insights from multiscale and multimodal data in microscopy and medical imaging, for example to image across physical scales (from molecules to cells, tissues and organisms) or to integrate anatomical, morphological, functional and molecular information to visualise and understand e.g. the heart, brain or whole organism.

[7] Including high profile investments such as Health Data Research UK and the Alan Turing Institute

- **Deep phenotyping through linking 'omics' to 'patient and population' data.** Linking molecular information on genomes, genes, proteins and metabolites to individual patient characteristics via health records, patient generated data and population cohorts in order to identify individual (patient) trajectories in disease risk, prevention, prognosis, or treatment-response in **precision medicine**; to move from standard care pathways to treating people with **multimorbidities**, e.g. by understanding common mechanisms.

- **OneHealth.** Taking a 360-degree approach to e.g. food safety, zoonotic diseases or anti-microbial resistance with integrated analyses of animal, human, societal and environmental health[8].

1.3 Timely opportunities for enhanced integration and sharing across a wealth of data assets

As biomedical research becomes increasingly data-rich, there is the potential to **accelerate and increase the efficiency** of research by improving the collection, curation, integration and availability of data for research (re-)use. The combination of regulatory strength and wealth in data assets provides a unique opportunity for the UK to safely and ethically integrate diverse data sets in novel ways and lead delivery of new data opportunities.

While significant progress has been made in some areas, this should now be expanded across methodologies and scales. To maximise the opportunities arising from the wealth of data, there is an **urgent need** to ensure data can be linked and shared in an agile and intuitive manner for broader interrogation by the research community.

In addition to providing new scientific insights, data integration provides opportunities for **new partnerships, including with industry**, which has interest in accessing, but also holds, valuable data assets which could benefit from integration. Supporting responsible collaboration and communicating the potential of existing and emerging capability of data science is essential.

1.4 Potential of data science for experimental design and conduct

Data science has the potential to transform the practice of science in many respects, and in many parts of the 'research system', particularly the design and conduct of research[9].

An area of unique opportunity in biomedical science is the **fusion of data derived models with empirical, scientifically derived models**, allowing the integration of real world (often sparse) data with fundamental biomedical insights that have been derived from experimental and theoretical science[10]. This offers the opportunity to fully utilise fundamental scientific knowledge, theories and equations alongside real-world data.

---

[8] https://www.who.int/news-room/q-a-detail/one-health

9 including automated evidence synthesis to guide commissioning, value of information approaches and fast access to linked data to guide adaptive designs, and automated throughput. These opportunities are also considered within the 2021 UKRI AI Statement of Opportunities: Transforming our world with AI

[10] For example, formalised as equations (e.g. Michaelis-Menten enzyme kinetics; Hodgkin-Huxley model for the electrical characteristic of excitable cells)

1.5 Improving biomedical data to improve research – a visual representation

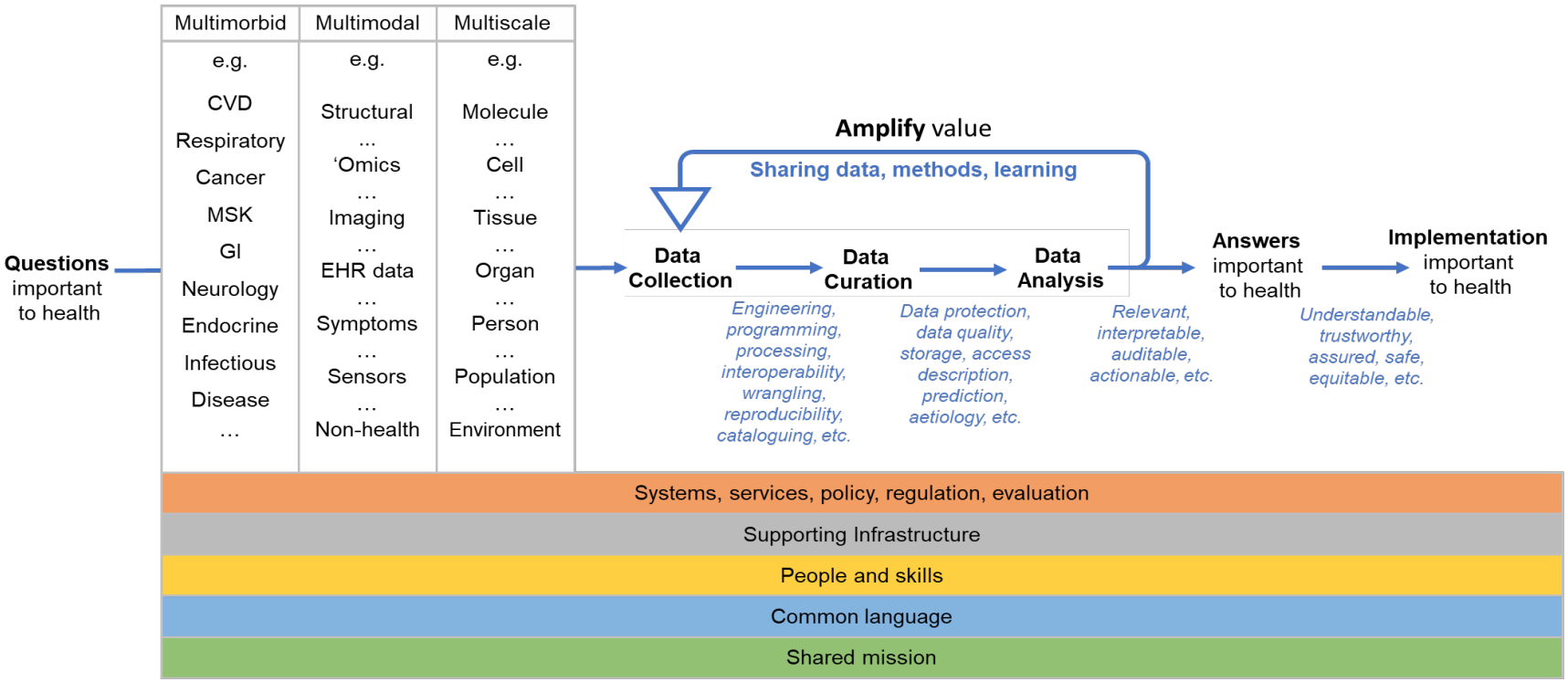


Figure 1 provides a visual representation of a pathway of robust biomedical data science.

Beginning with important biomedical or health questions progressing through data collection, curation and analysis to culminate in relevant, informative and actionable outcomes. **Data and methods sharing** within the pathway will produce better answers from the collected data and for future learning. Supporting communities in **understanding and developing methods** catalyses innovation, reduces fragmentation across the different applications of data science, enables the cohesive research approaches required and amplifies value. Each step within this pathway requires different areas of expertise to work together with **common language and aims** (including approaches to reproducibility, reusability and transparency).

2 Enablers of biomedical data science

In order to realise the significant benefits and opportunities in biomedical data science, there are a number of broad needs or enablers that will need to be addressed, with key themes including:

2.1 People, skills, diversity and equitable team science

Data science encompasses a wide range of expertise and capabilities (Figure 2), as does biomedical science. In biomedical data science there is a particular need to provide **cross-domain training throughout the career spectrum**, supporting training of 'wet laboratory' scientists who are data literate, 'dry laboratory' scientists who are biologically literate, and fully interdisciplinary scientists who combine or integrate both 'wet' and 'dry' approaches.

Wider levels of data and computational **literacy** would further increase the number of individuals with an awareness of what is possible, feasible and required and encourage a 'common language' to **support interdisciplinary teams**.

Successful biomedical data science teams seamlessly integrate multidisciplinary expertise and knowledge of biomedicine and data science. This **team science approach** enables co-design of projects and swift validation of model outputs in biological or clinical models. As a result, findings and solutions are more readily applied to real-world scenarios, with improved generalisation and reproducibility. However, team science approaches are currently not incentivised, or easily financed, through traditional response-mode project grants and fellowship schemes, exacerbating the risks of skills gaps within research teams (Supported by survey commissioned by MRC after review; data provided in figure 3 not seen by Review Advisory Group; Survey by Warwick Economics and Development).

To enable sustainable teams, without continual loss of data scientists to other well-paid sectors, an **equitable approach is essential**. All partners within a team should be valued and recognised appropriately, avoiding a 'service provision' model and disciplinary hierarchies.

Attracting talented individuals to biomedical data science also requires **visible, attractive career pathways** to reward and retain people who work in this field[11] and attract new talent, bringing additional capacity, skills and ideas. This has been a challenge where progression and promotion requirements are still focused on traditional publication metrics, although change is underway.

Enabling **cross-sector career transitions** across academia, medicine, and industry (including outside biomedicine) is particularly pertinent for data science, where talented individuals may see a better career offer from roles in industry. There also remains a shortage of **data scientists in visible academic leadership roles**. The UK would benefit from a strategic approach to enable more, and more visible, data science leaders, both for the benefit of biomedical research, and to inspire others to contribute to biomedical data science.

[11] An integral part of this is recognition of the importance and need for data engineers, curators, wranglers.

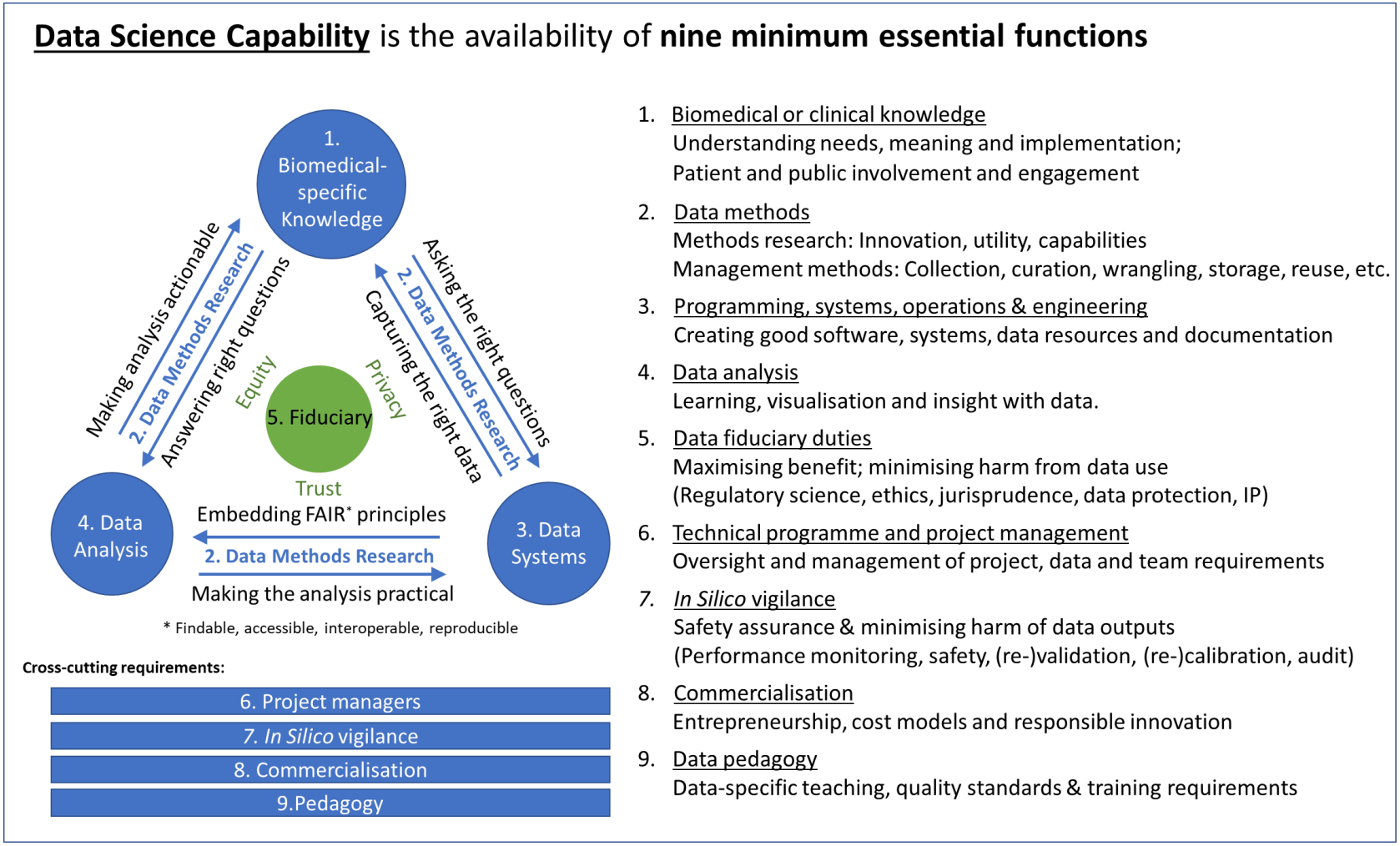
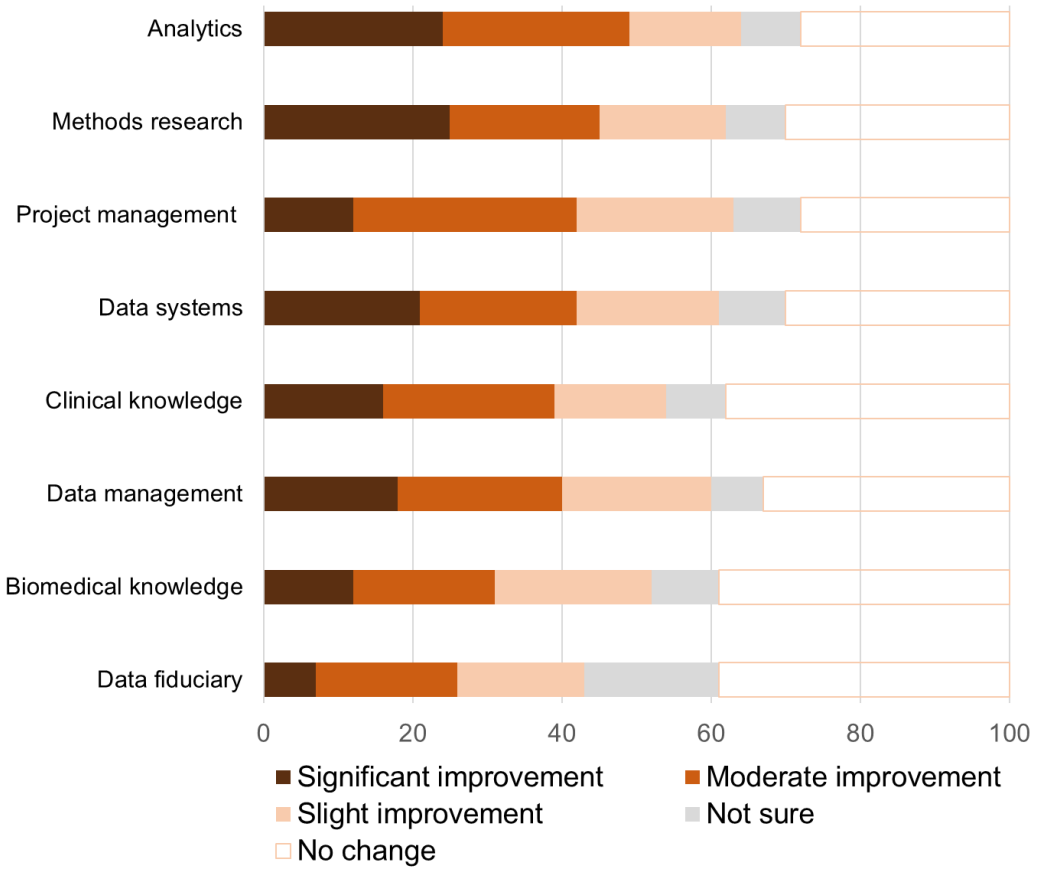


Figure 2 Different technical competencies, nomenclature and organisational structures will be required by different research settings. Instead, a shared framework of data functions is suggested as more portable guide to skills mobility and recognition across these different contexts.

The table provides an initial example of how this framework could be approached. Data-related functions are divided into those already embedded within biomedical research (*established* – variable levels of capacity and capability) and proposes other essential data-related functions that are currently recognised less consistently within biomedical research (*in development* - low to no capacity or capability).

Figure 3: Potential for additonal expertise within the research team to improve the delivery of current biomedical research projects



Percentage of 175 responses received from early-mid-career researchers actively engaged in biomedical research project in the UK; 2021 MRC-WECD Survey

Q: "Thinking about the expertise currently available within the project team, would additional expertise in the following capabilities improve the delivery of the project?"

2.2 Visibility, discoverability and accessibility of data and methodology.

The UK has a great wealth of data assets and maximising their use and integration is therefore a clear opportunity. However, the effort required to 'process' data so that it is **easily re-usable** (i.e. data engineering, curation, wrangling, linkage) should not be underestimated, representing about 80% of effort, compared to 20% for subsequent data analysis.

Building on and expanding the FAIR principles[12], as applied to data and methodology, would encourage **both data and methods** to be findable, accessible, interoperable and reusable. This requires training and support to manage, annotate, curate and **share existing data** in appropriate ways that enable broad data integration, where ethics and governance allow. In addition, **new data collection and consent** should occur in efficient and systematic ways to better facilitate automated management, usage, protection and sharing. Similarly, encouraging broad adoption of common **platforms for sharing of methodology**, such as code and user interfaces for interrogation of multiple data types, will allow for wider, more agile analysis across the community.

2.3 Increasing potential through digital research infrastructure

While this review focused on the extraction of insight and adding value to data, it also recognised the clear importance of digital research infrastructure, including that for data assets. It was acknowledged that there was a body of work being undertaken across UKRI regarding digital research infrastructure, to take forward the ambitions laid out in the 2019 'UKRI opportunities to grow our capability' [13].

The advisory group identified a major need to support **interoperability, federation, and linkage** of data from different sources, including the importance **of trusted research environments** (data safe havens) where this involves linkage to sensitive data such as personally identifiable information. Additionally, there is an emerging and accelerating role for **privacy enhancing technologies** that enable information sharing whilst retaining strong privacy guarantees. Here, there remain significant unmet capability needs, which will be essential for the research community and data donors alike, to have confidence that linkage will be feasible, in a timely manner, and with the **appropriate protections from harm**. Engagement with other stakeholders in this complex landscape is critical [14].

2.4 Requirements for innovation and methodological development.

There is a continued need for innovation and methodological development to further enable the interpretation and modelling of biomedical data. Statistical and biostatistical thinking remain core foundations in combination with subject knowledge, including for example in the appropriate treatment of missing data, to produce robust and reliable conclusions. Establishing **causality** is a fundamental challenge in biomedical data science, especially when bringing together data at scale, and there is a need for rigorous approaches that can deliver robust and accessible tools and methods for causal

[12] a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable (FAIR). Would also complement UKRI expectations for research data management and the Concordat on Open Research Data
[13] See footnote 6 on page 2
[14] Including HDR UK, NHS Transformation Directorate, NIHR and devolved nations' health research organisations, potential research users and data custodians and others.

inference at scale. The core steps to achieve this are highlighted in Figure 1.

Taking AI as an example, new theory and methodology is needed to properly and responsibly apply AI to biomedical research and draw causal inferences. The tools and methods developed in other sectors, e.g. language recognition, may not be directly applicable or transferable to biomedical research. In order to exploit AI tools for the use of high dimensional and messy health data for clinical benefit, without ignoring their limitations, there is a need to attract and support AI technologists to work in the biomedical domain. This includes a careful appreciation of the benefits and deficiencies of AI against more traditional statistical approaches, how AI can augment and be used in combination with traditional (robust) statistical methods, and the need for accurate uncertainty quantification and reproducibility of findings. This demands a community of researchers working seamlessly between AI and statistical domains, engaging with and included within real-world biomedical research groups.

2.5 Responsible use of biomedical data science.

Public and patient engagement and involvement in data science research is essential, informing the focus, design, delivery and implementation of research, as well as gaining wider public understanding and support for research using data. This assures that major biomedical research questions are addressed; and scientific, clinical and health impacts are realised.

3 Coordinated activity required to deliver the opportunity

The scale of the opportunity in biomedical data science necessitates an urgent and coordinated response across the research and funding space, so that the full benefit of data science is realised for the widest impact across biomedical and health research. There is a need for widespread culture change and clear communication of the potential so that ambitious and innovative data science approaches can be appropriately developed and supported.

3.1 Adapting response mode funding to encourage data re-use and team science

It should be made clear to the research community that applications incorporating data science research are specifically welcomed[15] and the **re-use of existing data** should be encouraged. Consistent **data sharing mandates** and more visible **collaborative approaches** across funders should further facilitate opportunities for data re-use and integration.

To ensure the merit of data science proposals is recognised, appropriate and sufficient **data science expertise** should be included in decision making wherever relevant. This could involve addition of expertise to existing decision-making committees (such as via a 'college of experts') or development of a new board or panel. Wherever possible, data science proposals should be encouraged via all routes, providing the widest possible access to support. However, a targeted initiative[16] of dedicated funding may be beneficial to provide a rapid and visible indication of interest.

Application processes should be reviewed to ensure that **interdisciplinary teams** are able to clearly demonstrate how they will combine the most appropriate data and methods to address the most important questions. All previous research contributions (including shared data resources, tools and methods) should be considered when assessing proposals and track record. **New approaches to evaluating impact** need to be developed to provide equivalent status to more traditional publication outputs.

In assessing the benefit of novel approaches, clinical feasibility and impact may be difficult to demonstrate. Two-phased funding, to demonstrate feasibility prior to development at scale, could be considered where the potential benefits outweigh the uncertainty. Funders should also consider the opportunity to **encourage team science best practice** within research organisations by linking evidence of recognition and career support to provision of funding.

[15] MRC's Research Boards all already feature a 'board opportunity' in data science, but this could be made more explicit to the community

[16] For example, a challenge-led call based on 'big questions' could also inform methodological gaps. Alternatively, a call focussed on data re-use would be a feasible and economical way to further demonstrate the potential of data science in biomedical research.

3.2 Attracting more data scientists into biomedicine

Data scientists and their skills are sought after in all sectors. Making informative, multi-modal, and interesting data **more easily discoverable and accessible** (see section 2.2) would enable and encourage data scientists to apply their skills to biomedical questions, rather than those of other sectors. This would be further strengthened by improved approaches to include data scientists within active biomedical research teams.

In addition to requirements for **attractive career paths** and an equitable approach with **appropriate recognition** to attract and retain individuals, there is a need for further visible support to **encourage interdisciplinary training** (see section 2.1) across the 'wet', 'dry' and 'fully interdisciplinary' spectrum of biomedical data science. Dedicated schemes could boost numbers but would require a local cross-departmental nurturing environment[17] to support individuals. Alternatively, interdisciplinary training could be aligned to new strategic investments (see below) to provide both support and visibility.

The need for increased data, statistical and computational **literacy** to increase awareness of the potential of data science and enable interdisciplinary teams who speak a **common language** could be supported through additional cross-domain training. This should be provided across the career spectrum, building on existing best practice within PhD training.

Finally, while there is some support for **cross-sector career transitions** through MRC and UKRI careers and skills activities[18], there are further opportunities to encourage and facilitate best practice and the option to embed targeted support in a visible investment to enhance opportunities to attract and retain data scientists.

3.3 Enhancing current national investments in biomedical data science

There are a number of leading UK large investments of relevance to biomedical data science[19] and the MRC Portfolio Review identified the need for strategic investments to provide national outreach and help to deliver the national agenda. Linking relevant large investments, potentially in themes, would support this ambition by **enhancing visibility and supporting collaboration**.

A broader **virtual network** could encompass both existing national investments and the wealth of regional or university-led investments[20] to enable the national spread of knowledge and activity. Networks could be based on challenge themes, combining data science with major biomedical research questions in a collaborative and co-led approach with modest additional funding.

[17] MRC's previous devolved Skills Development Fellowships provided such support, but would require a wider scale to provide the visibility and numbers required.
[18] such as CASE studentships and UKRI Innovation Scholarships.
[19] See footnote 5 on page 2 for details
[20] Such as data science institutes, data hubs, data holders and teams

3.4 A new 'flagship' national initiative in biomedical data science

While linking existing investments would increase visibility, it would not address all the areas identified within this review. An accessible and visible **'flagship' initiative in Biomedical Data Science** would provide the opportunity to deliver solutions to critical problems in medicine and health using data science while also complementing and partnering with existing national assets. Weaving multiple stakeholders, approaches and activities into a single flagship initiative would enable the step-change needed to:

- support the development of methodology and delivery of novel and exciting data science at scale, with data science included in biomedical domain centres of expertise;
- maximise the learning opportunities and wealth creation from UK biomedical data assets;
- enable access nationally to data science expertise and capabilities and leading data accessibility and re-use;
- provide an international landing point for UK biomedical data science;
- encourage more data scientists to apply their skills to biomedicine, supporting training across career stages, developing attractive career structures and enabling equitable team science;
- Encourage and facilitate team science across disciplines.

A **virtual geographically-dispersed alliance** could bridge across the UK's best biomedical expertise, data and strategy; simplifying recognition and messaging of the wide breadth of UK biomedical data science. The initiative will catalyse improvements in biomedical data practice and scale-up the quality and availability of biomedical data science nationally. This would enable the best people to tackle priority research questions, irrespective of geographical location. The changes in research culture it would initiate, especially regarding data accessibility/re-use and careers, would enable new and wider opportunities to attract diverse data scientists to biomedical research.

The advisory group were **strongly supportive of such a flagship initiative**, which would provide the necessary scale to position the UK in a leadership role for the rapidly developing biomedical data sciences.

Overall recommendations

The advisory group were unanimous that biomedical data science is both a substantial emerging opportunity, ranging across all relevant data sets[21], and that there is an **urgent opportunity and need to act**. As multiple sectors have already been transformed by data science, there is a real risk that biomedical research in the UK will be 'left behind' by developments in other countries. This does not have to happen. Given the right investment and support, the UK has the potential and infrastructure to be world-leading in biomedical data science. This in turn would enable the UK to return most benefit from industries investing in biomedical data science and health data assets.

Coordinated activity is required at all levels to deliver on the potential of biomedical data science, including (i) targeted interventions to improve research culture, appropriate mechanisms and incentives to enable data re-use and diverse-team-science led projects; (ii) providing interdisciplinary training across career stages and research contexts; (iii) creating attractive, accessible, transferable career structures; and (iii) enhancing visibility and flexibility of the UK's investments in data science.

There is an urgent unmet opportunity to add value to both biomedical scientists working with data and data scientists working in biomedical research, and to set a clear course and shared expectations to **support communities in developing better ways of working across multiple organisations**.

While progress is possible by adjusting and enhancing existing mechanisms and networking current investments in the short term, **a new visible flagship initiative would provide the opportunity to address all the major requirements** emerging from this review and position the UK in a global leadership role for the future.

[21] from molecules to humans, from animal models to ecosystems and spanning scales and modalities

Annex 1: Advisory group membership

Advisory group members

Cathie Sudlow (Chair); Professor of Neurology and Clinical Epidemiology, University of Edinburgh and Director of BHF Data Science Centre, Health Data Research UK

Rob Buckle; Medical Research Council, UK Research and Innovation; Chief Science Officer

Bianca De Stavola; University College London; Professor of Medical Statistics

Will Dixon; University of Manchester; Professor in Digital Epidemiology

Tania Dottorini; University of Nottingham; Associate Professor in Bioinformatics

Aldo Faisal; Imperial College London; Professor of AI and Neuroscience

Chris Holmes; University of Oxford, Alan Turing Institute; Professor of Biostatistics; Department of Statistics and the Nuffield Department of Clinical Medicine

Nicholas Jewell; LSHTM, UC Berkeley; Professor of biostatistics and statistics, public health

Marcus Kaiser; University of Nottingham; Professor of Neuroinformatics

Iain McInnes; University of Glasgow; Chair of Medicine; Professor of Rheumatology

Maddy Parsons; King's College London; Professor of Cell Biology, Director of Nikon Imaging Centre

David Westhead; University of Leeds; Professor of Bioinformatics

John Whittaker; GlaxoSmithKline; Vice President, Target Sciences, Glaxo Smith Kline Pharmaceuticals

MRC Head Office

Joanna Robinson, Head of Population and Systems Medicine

Richard Evans, Programme Manager

Claire Newland, Head of Data Science

Sam Rowley, Programme Manager

Nathan Richardson, Associate Director, Research Programmes Group

Annex 2: MRC Unit and Centre Portfolio Review - key relevant areas

Table 1: aspects of 'data' needs highlighted in each domain of the MRC Unit and Centre Portfolio Review

Domain	Key text	Long list, key areas of opportunity and need
Molecular and Cell	See 'Data' (Portfolio Review, page 12).	Data (page 13) • Need for standardization - not all working on our own models • Infrastructure – curation, storage, access, sharing, metadata • People – reward / collaboration, training, embedding in both biomedical and biostats, retention and career paths • How to support code development and sharing to support robust tools and reduce duplication? • Need for good user interface for the biologist • Analytics – quantitative, modelling, data visualization
Physiological Systems	See 'Cohorts and data' (Portfolio Review, page 19); 'Data and data analysis' (Portfolio Review, page 20).	Data (page 21) • Not just data science but also modelling • Critical to get EPSRC buy in as it is the Council that mathematicians really engage with. • Opportunity for interaction with Pharma (c.f. target validation) • Methodology not well supported in response mode or by career paths so could argue for strategic investment
Population and Public Health	See 'Data assets', (Portfolio Review, page 27); 'Data science' (Portfolio Review, page 30).	Improving data/information analytic capacity (page 29) • Enhancing the weight of the portfolio in methodological research and providing additional funding for method development • Building up human capacity in data science (integrating biological/systems insights with maths, data engineering, stats skills) • Retaining capacity despite the large pull to private sector (e.g. with more attractive salaries and research environments) • Developing new robust models for working with private sector data companies • Development of methods for testing interventions at scale and speed