UK-HDAN
Data Science for Health

UK-HDAN Research Roadmap

1 Overview

The UK Health Data Analytics Network (UK-HDAN) is an open community of researchers and practitioners who share an active interest in applications of 'big data' methods in health and social care. It brings together the health informatics and data science communities to exploit synergies, identify research challenges, and build new partnerships. As of January 2017 the Network has more than 450 members from over 100 HE, NHS and industry organisations.

Between January and March 2016 the Network organised a series of workshops, attended by over 160 individual participants, with the aim of developing a Health Data Analytics Research Roadmap. This document – a first draft of the Roadmap – provides a structured summary of the workshop outputs, with three components: an analysis of the *Healthcare Opportunities* – the ways in which the application of data science could transform health and social care; a summary of the *Data Science Research Challenges* – the methodological advances that will be necessary to realise the potential for transformation; and an outline *Ethical and Responsible Innovation* Framework, which lists non-technical issues that should be considered by researchers working in this field.

2 Healthcare Opportunities

The roadmap focuses on ways in which data-intensive methods, supported by data analytics could have a direct impact on the delivery of health and social care. There is also potential for indirect benefit through applications of data science in biomedical research, but that is out of scope – though there are areas of overlap. We identify five broad areas of opportunity, which share common assumptions about the availability and use of personal health and health-related data: *New Insights from Ubiquitous Data*, *Better Care Through Patient-specific Prediction*, *Personalised Care*, *New Models of Care*, and *Learning Health Systems*; they provide closely related but distinct views of potential for benefit.

2.1 New Insights from Ubiquitous Data

Fundamental to the transformational potential of data-intensive care is the opportunity to gain a more complete picture of individuals' health, lifestyle, exposures and experience than is currently available, by integrating and analysing data from multiple sources. The idea is to augment formal, intermittently acquired information, currently stored in health records, with data that provides a much richer, more continuous description of lived experience. Examples include data from mobile and wearable devices, environmental data, retail transactions, social media, patient-reported experience, digital footprint, utility usage and, more generally, internet of things. These data can be used both to underpin improved care for the individual and provide new insights at the population level. Potential broad impacts, illustrated with more specific examples in subsequent subsections, include the following.

Better informed care. Currently, healthcare decisions are based, typically, on limited information, sampled intermittently (and often unreliably) during formal interactions with the healthcare system. Data science has the potential to provide a more complete description of lived experience and health outcomes, to inform better-targeted care and to support self-care. This includes understanding severity and patterns of symptoms, triggers, behavioural determinants, and medication adherence.

Responsive care. Currently, healthcare is provided in a relatively standardised way, based on typical needs and responses to treatment. Data science has the potential to facilitate a more responsive approach to care, developing an understanding of what is normal for an individual, and detecting significant change. This can be used to influence behaviour, engage peer support, and target professional intervention at the point of need.

Population insights. Currently our understanding of disease and response to treatment 'in the wild' is limited. Data science has the potential to build a more complete understanding of the development, temporal characteristics, outcomes, and response to treatment of disease across populations – including their determinants and variability. This can be used as a basis for improving patient pathways and developing new treatments.

2.2 Better Care through Patient-Specific Prediction

Harnessing data analytics to make more effective use of pervasive data has the potential to improve care at the level of the individual by making patient-specific predictions – providing actionable information at the point of care. The idea is to develop patient-specific models that draw on both population and personal data to predict health outcomes – including prognosis and response to potential treatment – with the aim of deploying the right investigative, preventive or therapeutic intervention at the right time. Potential impacts include the following.

Precision medicine. Many conditions that have in the past been classified and treated as single diseases, are now recognised as clusters of diseases with similar symptoms but differing underlying pathology and, thus, response to treatment. Data science has the potential to discover new phenotypes (ideally endotypes), and allocate individuals to the most appropriate treatment group. Given the known influence on outcomes, of cultural and socio-economic factors, this approach should embrace behavioural as well as bio-pathology phenotypes.

Dynamic management. Existing patient pathways take a one-size-fits-all approach to condition management, militating against agile response to disease progression, comorbidities, changes in response to treatment or altered circumstances. Data-intensive methods have the potential to allow and support dynamic, individualised reconfiguration of pathways, modifying treatment and other interventions, and instigating investigations, in response to evidence of change in efficacy or risk.

Forestalling acute episodes. A large proportion of current NHS cost is associated with unplanned hospital admissions due to acute exacerbations of long-term conditions (eg respiratory, mental health), often accompanied by permanent reduction in quality of life. By building a detailed understanding of what is normal for a given patient, data analytics can be used to provide alerts for the individual, their family/friends and health care professionals, allowing early intervention to forestall acute episodes.

2.3 Personalised Care

Building on existing experience in the retail sector, data-intensive care has the potential to transform patients' experience of healthcare, providing services, information and advice relevant to their needs, empowering them to engage in their own care, and managing it in a way that suits their personal preferences. Different aspects of personalisation can potentially be subsumed in a virtual personal health assistant (eg mobile device app) that is fully aware of an individual's health status, treatment, preferences, history and context. Potential impacts include the following.

Patient choice. Current practice provides very limited opportunity for patients' personal preferences to influence care. Comprehensive deployment of data-intensive care presents an opportunity for patients to influence their care directly, both by providing explicitly stated preferences regarding treatments, personal goals, use of data, involvement of family and friends etc, and by inferring them from their interactions with personalised advice and feedback.

Experience sampling. Patient-reported experience of their condition is arguably the most relevant measure of impact on quality of life, but currently is recorded infrequently and unreliably – generally in face-to-face consultations – with little impact on the delivery of care. A data-intensive approach to care provides the opportunity to collect patient reported experience data more frequently/systematically and use it to influence care directly.

Personalised support. Currently, patients receive generic advice and very limited personalised feedback. Data-intensive care creates the opportunity to provide personalised feedback, explanation and advice directly relevant to an individual's condition(s), level of activation (commitment and skills to self-care), and context, improving their experience of the health system and empowering them to take greater control of their own health.

2.4 New Models of Care

There is a pressing need to develop new models of care, particularly for long-term conditions, delivering high-quality care in a community setting. Data-intensive methods have the potential to power new models of care in the community, increasing the focus on prevention, transforming care by supporting patients to manage their own health, and targeting resources intelligently to meet patient needs. Potential impacts include the following.

Wellness. Currently, healthcare resources are focussed mainly on treating individuals who are ill, despite the fact that much of the burden of disease is preventable. There is already a consumer market in wearable devices and associated data-driven methods to support and encourage healthy behaviours. Given financial pressures and skill shortages, it will become increasingly important for health and social care providers to extend this approach, building more holistic views of, and influencing positively, individuals' health behaviours.

Self- and collaborative-care. As the population ages and individuals live longer with chronic conditions, there will be increasing pressure for patients to take a more active role in manging their own long-term conditions, supported by relatively low-cost care in the community. Data-intensive methods have the potential to power this transformation by providing individuals and their carers with the information and tools to collaborate in producing and implementing their own care plans. For example, by identifying links between personal behaviours and health outcomes, data analytics can be used to drive behaviour-influencing prompts and personal decision-support tools, helping individuals to live independently for longer.

Targeted support. Currently, resources to support care in the community are spread thinly, with limited ability to respond to individual needs. Data-intensive methods have the potential to provide a much richer picture of individual needs, allowing more dynamic and targeted allocation of health and social care resources, providing tailored support/intervention where it is needed, whilst avoiding unnecessary deployment.

2.5 Learning Health Systems

Much of the emphasis in preceding sections has been on delivering better or more appropriate care for the individual. Data-intensive methods, drawing on system-wide data, also have an important role to play in the continuous characterisation and improvement of whole health systems, providing actionable information for policy-makers and service managers. Because there is considerable heterogeneity in population needs and responses (for example the same intervention may work in one population and not in another), it is important this takes a place-based approach, though some learning may be transferable across populations/systems. Potential impacts include the following.

Health surveillance. Currently, it is difficult to build a comprehensive and timely view of the changing health needs of a population, or to detect critical events. Data-intensive care creates the opportunity for systematic population surveillance, providing continuous information on population trends, incidence of infectious disease, unexpected drug side-effects, and environmental risks.

Realtime feedback. Currently, there is no way generally to study the operation of whole health systems in real-time; at best, partial snapshots are taken periodically. This makes it difficult to identify and evaluate opportunities for system-wide improvement. In a data-intensive care environment there is the potential to provide feedback on system performance in (near) real-time, at all levels of granularity (system-wide to individual GP practice), to inform continual improvement of structures, pathways, information flows, interventions and professional behaviours.

Targeting resources. It is currently difficult to find a rational basis for allocating resources within a health system, with changes to one component often creating unintended consequences elsewhere. The ability to study, and potentially model, whole-system operation in a data-intensive care setting, creates the opportunity to understand the full financial and workforce, as well as healthcare, implications of potential system reconfigurations, providing a rational basis for resource allocation.

3 Data Science Research Challenges

Driven by the work on Healthcare Opportunities, we have identified key data science research challenges that will need to be addressed if the full potential of data-intensive methods to transform health and social care is to be realised. A common thread is the need to 'industrialise' health data analytics, moving beyond the current 'cottage industry' approach to more scalable solutions. There are five broad themes: *Integrating Heterogeneous Data*, *Dealing with Imperfect Data*, *Predictive Models*, *Identifying Subgroups*, and *Human-centric Systems.*

3.1 Integrating Heterogeneous Data

New healthcare opportunities bring extreme challenges in combining disparate kinds of data of varying reliability, at volume. The need is for new methods of representing heterogeneous data and automating it's integration. Specific challenges include.

Representation and data models. The heterogeneous nature, reliability and sampling of health and health-related data presents significant problems for data modelling and representation. There is a need to develop sophisticated approaches (ideally standards) for capturing the complexity of data, including dealing with imprecision, particularly in temporal relationships.

Provenance. When data from multiple sources, of varying reliability, is combined, it is important to keep track of the provenance of derived data and inferences so their reliability can be ascertained. This requires a semantics for provenance and machinery for propagating properties – particularly reliability and permissions – through complex chains of use and reuse.

Robust linkage. Linking data by individual, location, condition etc is fundamental to the power of data analytics. With data from diverse sources, this is, however, not always straightforward. Handcrafted tools can be created to deal with some situations, but the approach is not scalable. There is a need for probabilistic and machine learning based linkage methods to provide robust automated solutions.

Using context. Context can be critical to the correct interpretation of data, for (a trivial) example, the amount of physical activity undertaken may depend on the weather. In part, this can be thought of as a special case of the linkage problem, when contextual data is available, but there is an interesting and important challenge in inferring (latent) context from data.

3.2 Dealing with Imperfect Data

Health and health-related data 'in the wild' presents challenges in the form of heterogeneous sampling, missing or corrupted values, and temporal drift. These problems are exacerbated by the use of new forms of data, often collected under uncontrolled conditions. It is unrealistic to imagine that controls could be put in place to ensure perfect data, or that manual 'curation of data is a scalable solution. Rather, it is important to be able to deal with the data as it comes. Specific challenges include.

Heterogeneously sampled data. The fact that health and health-related data are typically sampled irregularly, over a wide range of frequencies, and rarely simultaneously presents a fundamental problem for many data analytic methods, given the generally time-varying nature of the data. There is a pressing need to develop principled approaches to working with such data.

Missing and anomalous data. It is characteristic of health and health-related data that values are often missing or corrupted. Where simple models of data generation apply, there are well-established methods for imputing missing data and detecting (and replacing) anomalous data. This is not, however, a solved problem for more complex (eg timecourse) data, where new methods are required.

Temporal drift. Due to a mixture of instrumental and human factors, data that should remain constant over time can drift. In order to draw correct inferences it is clearly important to take this drift into account. This is a challenging problem, unless calibration data is available (generally it is not), and there is a need for principled methods for inferring drift directly from population data.

Integrating inference and data cleaning. Although the problems of heterogeneous sampling, missing and anomalous data, and temporal drift can be considered independently as precursors to drawing inferences from the data, a more elegant approach is to treat data 'cleaning' as an integral part of the inference problem. There is significant scope for developing such approaches.

Managing uncertainty. Health and health-related data generally carry some degree of uncertainty, whilst further uncertainty is introduced by inference mechanisms. It is important to develop methods that keep track of uncertainty, so it is clear what degree of reliance can be placed on results.

3.3 Predictive Models

The ability to make predictions on the basis of observations, and detect deviations from what is normal for a given individual is key to the vision of data-intensive care – and one of the most significant challenges. At its heart is the idea of n of 1 analysis – comparing the state of an individual to their own past states, rather than a population norm. Specific challenges include.

Holistic models of individuals. The ability to make patient-specific predictions and detect deviations from normality is fundamental to many healthcare opportunities. This requires building an holistic predictive model of each individual, drawing, ideally, on everything in their history, plus general knowledge of physiology and pathology. This is a long-term goal, but advances in this area will be critical to the success of data-intensive care.

Modelling complex temporal behaviour. A particularly challenging but important sub-problem is that of modelling complex temporal behaviour. Biological functions are typically episodic, semi-repetitive over many timescales, and thus difficult to model. For example, what would it mean to say that the data from the accelerometer in my wearable was significantly different today from yesterday?

Borrowing strength. Although predictions based on patient-specific models are the ideal, there may be insufficient historical data to make confident predictions under some circumstances (eg a novel situation for the individual). To deal with such situations it is important to develop methods to draw strength from the population of individual models (people like you).

3.4 Identifying Subgroups

Discovering structure within data is key to many of the healthcare opportunities afforded by data-intensive care. Examples include discovering subgroups of susceptibility, disease and response to treatment, where differing biological mechanisms may be at play. Although this is a relatively mature area of research, healthcare data have particularly challenging characteristics, and the consequences of creating spurious groupings or allocating individuals to the wrong group are potentially serious. Specific challenges include.

Robust identification. The high-dimensional, unreliable, heterogeneous and heterogeneously sampled nature of health and health-related data, together with the presence of confounding (possibly unobservable) variation, pose a particular problem for robust subgroup identification. There is a need to develop more robust methods that also provide causal explanations of structure.

Feature selection and dimensionality reduction. A common approach to simplifying the grouping problem is to reduce dimensionality, either by selecting features particularly relevant to classification, or by extracting latent variables that define a subspace. Again, the characteristics of health and health-related data make this particularly challenging, and there is a need for new methods.

Transferability. An important issue in scaling data-intensive methods is the transfer of learning between populations. Given differences in genetic makeup, disease prevalence, socio-economic profile and culture between populations, this often proves non-trivial – for example, features that prove valuable for grouping in one population may not in another. There is a pressing need to develop principled methods for transferring learning between populations.

3.5 Human-centric Systems

New data-centric approaches to healthcare pose serious challenges in usability. There are serious risks of information overload and limited communication bandwidth, which require intelligent approaches to the human-machine interface. This is not just a matter of good design, but rather of building in new forms of intelligence. Specific challenges include.

Natural communication. The ability to extract meaning automatically from natural communication – speech, unstructured text, images – is fundamental to creating human-centric systems for patients and healthcare professionals. Although these are relatively mature areas of research, significant advances are still required.

Adaptive interfaces. Effective data-intensive systems will rely heavily on the ability of interfaces to adapt to both the user and their context, using knowledge of past interactions and experience to filter information for relevance, support peer-peer and patient-professional collaboration, and provide persuasive prompts. This is still an under-explored area of research.

Decision support. Ultimately, the purpose of data-centric systems is to support users – patients, professionals, individuals, teams – in making decisions about care. This requires methods of presenting actionable information in ways that are intuitive (including visualisation), with explanations that are relevant to the user. This is also an under-developed area.

4 Ethical and Responsible Innovation Framework

Drawing on input from the UK-HDAN workshops, and patients (mainly in the area of mental health), we have begun to develop an Ethical and Responsible Innovation Framework for health data analytics research. The framework is currently less complete than the other elements of the roadmap, but we are keen to prompt debate and receive feedback. The issues we have identified so far, and which should inform researchers are as follows.

Privacy and confidentiality. There are clear issues of privacy and confidentiality for patients and their carers in collecting, sharing and linking real-time data.

Consent and consent management. Individuals should be able to control the ways in which their data are used, and should be able to change their preferences dynamically.

Co-development. Patients, their carers and healthcare professionals should be involved directly in developing technology interventions which will generate, or rely on, personal data, to ensure that they meet real needs and are acceptable to users.

Choice and personalisation. Patients should be offered data-driven interventions as a matter of choice (rather than prescription), and should be able to personalise solutions to meet their own needs.

Empowerment and control. Data-driven interventions may empower patients to take more control, shifting the balance in the doctor-patient relationship but there is a risk that they could also be used to impose boundaries rather than removing them.

Impact on relationships. Patients may be concerned that data-driven interventions could impact on human relationships, substituting for personal interaction and reducing tolerance of diversity.

Fair access and support. Data-driven approaches to health technology could favour patients with higher levels of digital awareness and greater access to technology, leading to increasing health inequalities.

Perceived unfairness. Although there may be good reasons to provide different interventions for different patients with superficially similar conditions, there is a challenge in communicating this.

Information vs decisions. The human in the loop is essential. Patients, carers and healthcare professionals should be provided with relevant information, but they should make the decisions.

Explanation. It is important for users to be provided with accessible explanations of the basis for inferences that have been made and may affect their care decisions.