University of Manchester: Research Programme

Overview
In the first two years, our programme of research will focus primarily on health data analytics (HDA), exploiting the University’s unique strengths and partnerships to accelerate the Turing’s contribution to the Health application domain, whilst driving fundamental advances in data science. To maximise the potential for impact, we will work with Turing affiliates, and the wider HDA community, to build a coherent programme of collaborative research and seek new funding partners. We will also contribute to other Turing programmes, particularly those at the interface between Health and Urban Analytics.

Rationale
The potential for data science to transform health and social care is widely recognised. Opportunities include: delivering better care using patient-specific prediction, supporting self-management and collaborative care, gaining new insights using non-traditional (eg wearable, IoT) data, and dynamically optimising healthcare systems. Although there will be ‘easy wins’, realising the full potential will require fundamental advances in data science methodology. Data science challenges include: integrating heterogeneous data sources, robust inference from imperfect data, characterising complex temporal structure, identifying functional subgroups, and the marriage of human and machine. Progress will depend on establishing close relationships between health domain experts and data scientists, ensuring existing methods are applied optimally to real-world problems, whilst exposing fundamental data science challenges to methodology researchers.

Potential for Impact
For the University, the opportunity to work closely with colleagues in the Turing will enhance significantly an existing area of research strength, and our ability to deliver real benefit to healthcare partners. For the Turing, we will bring new energy to an important application domain, whilst exposing and addressing fundamental data science challenges. For the wider data science landscape, we will engage with the HAD community in the UK and globally to disseminate the outputs of the Turing programme and ensure relevance.

Outline Programme
Our proposed programme of research will address fundamental data science challenges in the context of Health applications. It will draw on expertise from, and contribute to, most of the Turing strategic research themes (architecture/algorithms, security & robustness, ML & AI, complex structure, humans in a connected world, ethics). Our aim is to ground the research by using real healthcare data to address real needs, but with a focus on reusable tools and methodology. The challenges are open-ended, but we expect to make significant progress in the first two years, establishing strong teams and collaborations, developing a deeper understanding of the challenges, and producing promising initial results.

In **months 1-6** we will work with healthcare and civic partners, existing Turing Fellows, and the broader HDA community, to understand domain issues, analyse preliminary data and identify suitable case studies, via a series of sandpit and workshop events.

In **months 7-24** we will undertake substantive methodological work, using the selected case studies, and interacting closely with existing Turing research themes. We will also work with Turing partners (academic and external), and the broader HDA community, to develop for the Turing a comprehensive and coherent programme of Health data science research, building on our core programme and the road-mapping work already undertaken by UK-HDAN.
In months 19-24 we will seek opportunities to exploit the preliminary results of the research in a healthcare setting, working with public, charity and industry funders (over 320 companies in our existing health data networks). We will also work with these funders to secure long-term funding for the continuing programme of fundamental research.

1. **Predictive healthcare**
Healthcare systems should move away from being reactive and focused on illness, and embrace approaches that are proactive, preserve health, and prevent illness. This requires the ability to predict illness and disease progression in advance; some clinical decisions are already guided by predictive models (e.g. QRisk for 10-year risk of myocardial infarction or stroke). We will develop models to predict expected health outcomes over time, incorporating uncertainty, at the individual and population levels. Individual-level prediction models should draw on both population and personal data to better predict health outcomes and deploy the right investigative, preventive or therapeutic intervention at the right time.

**WP1.1 Dealing with complex data.** Integrating data from multiple unevenly sampled sources (e.g. from sensors and EHRs) with non-random-missingness (e.g. data from only those patients visiting clinic) and competing risks due to alternative outcomes.

**WP1.2 Learning individual-level prediction models.** Learning from repeated events at an individual level and providing personalised predictions (for e.g. hospital readmission) through combining past outcomes for an individual with population data; exploiting learned heterogeneity/subgroups in risk prediction (linked with subgroup discovery below); individual uncertainty modelling taking into account missing covariates and heterogeneity in data quality/completeness.

2. **Patient subgroup discovery**
Individuals can vary markedly in their response to treatment, for largely unknown reasons. However, current medical practice is still largely based on evidence for the average response. We will identify subgroups who will respond well to allow targeted treatment regimes with better cost-effectiveness ratios. Subgroup identification is also important for translational bioinformatics, linking phenotype to genotype, potentially providing a deeper understanding of biologic treatment pathways.

**WP2.1 Improving subgroup identification.** Clustering complex longitudinal multi-dimensional data with varying measurement frequency and informative drop-out (e.g. EHR data); selecting relevant features; using subgroups to differentially predict response to treatment; joint predictive modelling and subgroup inference.

**WP2.2 Assessing subgroup structures.** Assessing the quality and robustness of clusters in the absence of ground truth; power/sample-size calculations for subgroup identification.

**WP2.3 Transferability.** Transferring information from identified subgroups between populations given known differences in genetics, disease prevalence, socio-economic profile and culture.

3. **Human-centred systems**
Ultimately, data-centric systems must support human users (patients, professionals, families, teams) in making decisions. New data-centric health technologies pose serious challenges in usability. Despite intensive research, further advances are required in natural communication (speech, unstructured text and images). Adaptive and interactive interfaces can provide improved communication efficiency. Decision support requires an intuitive, relevant and accurate presentation of actionable information to human users with different levels of knowledge and expertise.
WP3.1 Understanding the healthcare data lifecycle. Empirically examining the relationship between human behaviour and data in healthcare systems. For example, understanding ‘data journeys’ in the NHS: how to minimise error, ensure provenance and facilitate interpretation when moving from lab to genetic technologist to clinical bioinformatician to clinical geneticist to clinician to genetic counsellor to patient to carer.

WP3.2 Shared decision making to support personalised care pathways. Methods to support decision making based on time-varying multi-modal data combining regularly sampled high-resolution data and sporadic low-resolution patient feedback (e.g. from patient/citizen sensors); designing adaptive interfaces for shared decision making, considering how the same data/decisions are presented and combined in different contexts; decision analytics with uncertainty, preferences, bias and data evolution over time.

WP3.3 Data-intensive self-management using digital biomarkers. Personalised learning systems for identification of digital and behavioural biomarkers (in combination with clinical biomarkers) for early detection or monitoring of disease; support for self-care using multi-modal data (wellbeing sensors, voice/text analytics, social activities, diaries, sleep patterns, peer support) to detect behavioural change.

4. Cross-cutting workpackages
The above projects will be further enhanced through cross-cutting workpackages, addressing issues common to Health applications.

WP4.1 Dynamic consent, disclosure risk assessment and synthetic data production. Drawing out and solving privacy problems to enable successful deployment of HDA technologies.

WP4.2 Optimal design. Developing methods to identify which potentially costly streams of data a health system should bring online.

5. Contributions to other programmes
In addition to the directed projects outlined above, we will contribute to existing and future Turing projects, particularly in urban analytics and social science, and other areas where we have expertise that complements that of existing partners.