Proposal for an Alan Turing Institute Partnership Project

Project title: Knowledge Discovery from Health Use Data (KNIFE)

Investigators: **Dr William Marsh, Prof Norman Fenton, Prof. Martin Neil** (EECS, QMUL)
Dr John Robson (Blizard Institute, SMD, QMUL),

1 Introduction and Motivation

The increasing use and capability of Electronic Health Record (EHR) systems has made available large databases of patient records, linked across different health providers. These databases contain information about patients' use of the different health services, treatments and prescriptions. The data, collected for clinical management or financial reporting, has many actual and potential uses including discovering causes, optimising health delivery, allocating resources and choosing treatment. However, there are both practical and technical challenges to overcome before these benefits can be achieved.

The overall objective of the project is to lay the foundations for a transformative approach to patient-linked health data, making it accessible for both medical and data science researchers to fully exploit. We will achieve this by working with two groups who are custodians of data of this type in East London

The data is both large and complex covering over a million patients in East London. The data is relational, with multiple entities (e.g. patient, disease, clinical service). This is immediately a challenge given that familiar statistical techniques assume 'flat' data (or hierarchical clustering). Moreover, the entities are individually complex and variable (e.g. different data for different patients). Time is important: some diseases are resolved in a single episode, perhaps involving multiple interactions with the health system, but many others are chronic, with time characteristics (e.g. time to diagnosis, frequency of admission to hospital) of potential interest to health researchers. From these characteristics, we identify the areas of challenge that this data presents.

Challenge of Understanding the Data and It Potential Use. The data arise from the operation of the health services, so the understanding of the data contents is embedded in the health community and not accessible to the wider AI and ML communities. Although the data are held in relational databases, we hypothesise that the relational schema only captures semantic relationships to a limited extent. For example, a typical EHR uses a hierarchy of forms: common data at the top with a choice of lower-level forms to enter the data considered relevant. As a result, considered just as attributes of a patient, much of the possible data will be absent though not actually 'missing'. In a similar way, the understanding of the potential uses of the data is also well known to the health community but not mapped to the machine learning or statistical techniques needed.

Challenge of Knowledge Elicitation and Modelling. Achieving the full potential from the data requires knowledge of health care processes. Aspects of this may need to be modelled as part of the data analysis: for example, analysing the effect of possible changes in medical practice requires causal models and causal relationships cannot be distinguished from data. It is understood how to build causal models at a smaller scale but it is a new challenge to build knowledge-based models capable of working with data of this size and complexity. Additional benefits of a knowledge-based approach include examples such as: a) the accuracy of the data likely relates to how it was recorded – at the time of consultation or separately; b) medical care evolve over time (e.g. new guidelines are introduced) which will create systematic changes in the data; c) information (e.g. interventions and outcomes) may be latent, needing to be inferred.

Statistical Modelling Challenge. Existing approaches to analysis of this data [e.g. 1, 2] often use conventional statistical techniques by extracting a 'flat' dataset from the relational database, relevant to the problem of interest. This approach remains relevant. However, advances in Statistical Relational Learning (SRL) [surveyed in 3, 4] offer the potential to move beyond 'conventional' models, but with challenges of both scale and applicability: which of the wide range of potentially applicable models would answer the questions of interest? For example, much emphasis in the literature is on predicting the existence of links between entities, but it is unclear how important this is for EHR data.

Challenge of Efficient and Acceptable Data Handling. Existing studies using data from EHR systems, require extensive 'data wrangling' to extract and cleanse a usable dataset. This work is largely manual and very time consuming. Challenges include finding relevant data, coding data and missing data but also the more conceptually straightforward process of repeatedly writing, running and checking queries. Acceptable ways to host the data, especially if more complex ML techniques are to be used, is also a challenge. Although QMUL SMD has access to computing resources with appropriate security (e.g. the recent HDR award), the information governance issues are not straightforward.

These challenges distinguish this data from the *most common* examples used in large-scale data mining. Some applications of relational machine learning (e.g. social networks) are primarily concerned with a single entity or are familiar within the CS domain (e.g. authorship and citation of papers, web pages). Work on medical data is inter-disciplinary, and it is challenging to to bring together the relevant research communities, establishing both communication and trust.

2 Project Team and Expertise

Dr William Marsh (EECS, PI), is a member of the Risk and Information Management (RIM) research group (led by Fenton) and has played a major role in RIM's previous work on decision support for medical applications, in trauma, forensic psychiatry, musculo-skeletal injury and in the PamBayesian project, which has case studies in Rheumatoid Arthritis and Diabetes. A major component of this work is the elicitation and development of causal models using the necessary combination of data and expert judgement (notably concerning missing causal factors) – these models are implemented as Bayesian networks (BNs)

Prof Martin Neil (EECS, CI), has led much of the BN algorithm development work of the RIM group, including the breakthrough dynamic discretisation algorithm. He has been PI in projects to the value of over £4 million and has published one book and 80 refereed articles. He is a Director of Agena Ltd.

Prof Norman Fenton (EECS, CI), is Director of the RIM group and formally a Director of Agena Ltd. The RIM group is world-leading in its practical applications of Bayesian modelling and risk quantification in diverse areas including medicine, operational risk in finance, security and defence, legal and forensic arguments, systems reliability, transport safety, software project risk and football prediction.

John Robson (SMD, CI) is a GP and clinical lead for the Clinical Effectiveness Group (CEG) based at QMUL, as well as a cardiovascular lead for Tower Hamlets PCT. He is an author of QRisk (cardiovascular risk estimation) and QDScore (risk of developing type 2 diabetes). He was chair of the NICE guideline on lipid modification and CVD risk estimation 2008-10.

2.1 Clinical Data Groups

The work will be done in collaboration with two clinical data groups. All data will remain wholly under their control and access will be subject to specific agreements in each case. However, given the objectives of the project, direct access to data is not essential.

The Clinical Effectiveness Group (CEG) (https://www.qmul.ac.uk/blizard/ceg/)
The CEG (of which Robson is a member) is a partner in the Discovery East London programme, involving GPs, acute/mental health and the Clinical Commissioning Groups (CCGs) from across the inner north east London Boroughs. The programme will establish a secure data service with linked combined identifiable data for all direct health care in east London, accessed through an API. At present, data from GPs is included; work to link hospital data is in progress. This group is particularly interested in improved automation for data wrangling.

Tower Hamlets CCG (See Letter of Support)
The TH CCG has assembled a pseudonymised patient-linked dataset of health use episodes, for service improvement. This work is independent of the Discovery programme; although data originates from the same sources (GPs and hospitals) it has been extracted to a DB to which CCG staff have direct access.

Commercial Collaborators
There is also a possibility of collaboration with an SME, providing health monitoring. The origin of the data differs, so it might provide an interesting contrast but is not essential to the project.

2.2 Collaboration with the Turing Institute
Initial discussions will be held with the Professor Chris Holmes, who we understand will lead the soon-to-be -announced Turing Institute Healthcare Programme, to identify potential collaborators. These include Professor Mihaela van der Schaar, University of Oxford; Dr Charles Sutton, University of Edinburgh; members of the Department of Statistics, University of Warwick and the Probabilistic Reasoning group at UCL, with whom we share a common focus on Bayesian Networks. So far, relevant Turing project we have identified are AIDA (AI to automate the data analytics process) and PDQ (a library developed for generating database query plans over multiple data sources).

3 Programme of Work

3.1 Detailed Objectives
The overall objective of the project can be achieved with the following detailed objectives:

1. Understand the data and type of questions that health researchers wish to answer and present this information in a way that is useful to the AI and ML community.
2. Identify what AI/ML methods could reasonably be used to address the questions of medical interest and understand the need for additional causal knowledge for effective data analysis.
3. Create a generative model to enable data of a similar shape to be generated free of data governance concerns.
4. Understand the practical difficulties of data wrangling and the potential for greater automation.
5. Prepare proposals for future work.

3.2 Tasks and Output
The project will be made up of the following tasks. We will not necessary apply all tasks to all the datasets nor necessary cover all details of any dataset.

Task 1: Data Description at Several Levels of Abstraction.

- Create a simplified model of the data generation process, covering the organisations involved, using a mix of text and diagrams.
- Create a semantic schema, using web ontology techniques or related techniques [4]. We will also investigate the use and role of the medical ontologies used to tag data (primarily SNOMED).
- Investigate the mapping between the semantic schema and the storage schema, its use for data access and the potential for automating its construction.

This task will result in a description of the data that can be shared with potential collaborators. A complete data dictionary may not be necessary if the same concepts recur in different parts of the data.

Task 2: Generative Model. Building on the work of task 1, this task will:

- Develop a sequence of generative models, based on the description of the data generation process, forming the most detailed 'executable' description of the data.
- Use the models to generate synthetic datasets and, if possible, compare the descriptive statistics, using this to refine the generative model.

This task will result in a 'data sandpit' to allow other researchers to evaluate analytical and learning algorithms. The true values of many of the parameters used in the generator will be unknown, so (even if the structure is correct) the synthetic data will not fully correspond to the actual data. However, researchers will be able to modify and rerun the generator, creating a test bed for algorithms.

Task 3: Data Use, Knowledge Elicitation and Analysis Methods.

- Elicit from data custodians the uses of the data and categorise the data analytic challenges. We will consider both what is currently possible and what might be achieved in future.
- Review available statistical / ML (including SRL) techniques against the data analytic requirements.

- Investigate how causal knowledge could be elicited and modelled for data of this complexity (with examples, if possible).
- If possible, carry out 'proof of concepts' analyses, using either real or synthetic data (Task 2).

The task will result in a clearer definition in the data analysis and knowledge modelling research needs, distinguishing these from what can be achieved with current techniques.

Task 4: Practical Data Handling issues.

- Use the results of task 1 to investigate the data quality issues, including clarifying the concept of missing data given the clinical discretion in choosing which values to record.
- Catalogue the range of data wrangling issues encountered in practical and determine the type of knowledge needed to provide automatic assistance.

This task will result in a clear set of challenges for improving the efficiency of data analysis, using current analysis techniques.

Task 5: Outreach and consortium building. This task will be to build a consortium of researcher to collaborate on future work. Guided by the leadership of the Turing Health Programme, we will identify and contact potential collaborators. Workshops (or a workshop) will be run to share ideas and we expect to prepare follow-up proposals.

4 Future Work and Impact

The final stage of the project will be to build consortia to collaborate on future research. Although uncertain at the moment, possibilities include:

- Automated aids for data cleansing and extraction
- Tools for schema discovery and query translation to provide sematic access to data
- Refining data modelling techniques and, in collaboration with clinical groups, examples of data analysis.

The impact of the proposed project will be to make this possible by building a network of researchers interested in knowledge discovery from linked health records, able to communicate across disciplinary divides. Both the descriptive and generative data models will both allow potential collaborators access to the research challenges and provide techniques that can be applied to related projects.

5 Resources Requested

The total budget is £103,850, for one year. Marsh (40%) will lead the discussions with medical groups and potential collaborators and the data modelling work (Tasks, 1, 2 & 4); Robson (8%) will provide clinical guidance, being the main source of information about the data generation processes and the data contents; Neil (5%) will contribute to Task 3 and Fenton (5%) will assist with consortium building (Task 5). A junior RA (100%) will assist with the detailed work, working alongside the medical groups. Resources are also requested for travel (£1k), to visit potential collaborators and to run one or two workshops (at the Turing Institute if possible). A total attendance of 50 people has been allowed for, at £40 each. A full costing is available on QMUL's worktribe system (project ID 413544).

The project will take 12 months, starting between 1 September 2018 and 1st January 2019.

6 References

1. Robson J., Dostal I., Madurasinghe V., Sheikh A., Hull S., Boomla K., Griffiths C., and Eldridge S. NHS Health Check comorbidity and management: An observational matched study in primary care. British Journal of General Practice (2017) 67(655) e86-e93
2. Homer K, Boomla K, Hull S, Dostal I, Mathur R, Robson J. "Statin prescribing for primary prevention of cardiovascular disease: a cross-sectional, observational study". The British Journal of General Practice. 2015;65(637):e538-e544. doi:10.3399/bjgp15X686113.
3. M. Nickel, K. Murphy, V. Tresp and E. Gabrilovich, "A Review of Relational Machine Learning for Knowledge Graphs", in Proceedings of the IEEE, vol. 104, no. 1, pp. 11-33, Jan. 2016. doi: 10.1109/JPROC.2015.2483592
4. Sun, Yizhou and Jiawei Han. "Mining heterogeneous information networks: a structural analysis approach". SIGKDD Explorations 14 (2012): 20-28.