R13

Hub CTSU	Host University: University of Oxford
Supervisor Michael Lay michael.lay@ctsu.ox.ac.uk	Co-supervisors: Jim Davies
Is the project clinical or non clinical? Non clinical	
Title of PhD project? Exploiting Unstructured Data in Clinical Trial Settings	

Background to the project: Important information about diagnosis, treatment, and outcomes is often available only in the form of unstructured data: in clinical or laboratory reports, in patient notes, or as free text responses on case report forms. Even where the information exists also in coded form (e.g adverse event reports in randomized trials), there may be questions as to the accuracy or completeness of the coding. Purely manual interpretation of this unstructured data is costly and error-prone; biases may be introduced that are difficult to detect or correct for; the process cannot simply be repeated under new assumptions. For these reasons, efforts are underway to apply natural language processing technologies to the automatic or semi-automatic interpretation of unstructured data. An approach in which these technologies are applied as part of a fully formal data management process – audited, repeatable, and based upon a comprehensive treatment of data definitions and data transformations as linked, versioned metadata – would reduce the cost of re-use of unstructured data for research and other purposes, and improve the quality of any subsequent analysis. The Clinical Trial Service Unit (CTSU) at Oxford acquires large quantities of unstructured data from its large trials, UK Biobank and other high-value, long-running cohort studies, such as medication lists and pathology reports, and has considerable experience in trial design and delivery. The Big Data Institute provides expertise in computer science, statistics, and data engineering, together with an ideal environment for the investigation of health data research questions.

What the studentship will encompass: The studentship will be focussed upon the development and evaluation of methodologies for the management and transformation of unstructured data collected as part of randomized trials. This will involve: a systematic review of literature in natural language processing, domain-specific modelling, model-driven transformation, data governance, trials design and compliance; the design and implementation of techniques for automatic analysis, de-identification, and quality assurance; the development of metrics for measuring the applicability of these techniques to different classes of unstructured data; the development of domain-specific modelling languages and ontologies for the classification and management of information contained within and derived from unstructured data; the establishment of key properties of these languages and ontologies, in terms of mathematical foundations and relationships to alternative approaches. The HPS2-THRIVE and HPS3-REVEAL trials constitute a valuable resource for the development and evaluation of techniques: not only have these trials collected large quantities of unstructured data, including more than 42,000 medication reports, but this data has been manually interpreted against an agreed ontology – at considerable expense in terms of time and clinical effort; the raw, unstructured data and the coded interpretation will be made available to support this research.

Detail of supervision, including the roles of any named co-supervisors: *Dr Michael Lay*, Head of Project Information Science, Nuffield Department of Population Health (lead supervisor, clinical trials design and compliance, data governance, analysis, and quality assurance) and *Professor Jim Davies*, Professor of Software Engineering, Department of Computer Science, and Director of Clinical Informatics, Oxford NIHR Biomedical Research Centre (natural language processing, domain-specific modelling, model-driven transformation). Dr Lay works on the THRIVE, REVEAL and UK Biobank projects.

Detail of any planned field work/ Secondments/industry placement: No field work is required, although opportunities for collaboration will be available through the MRC Hub Network, the Farr Institute, UK Healthcare Text Analytics Research Network, the Oxford Big Data Institute, and existing research collaborations, including Stanford University (NIH National Center for Biomedical Ontology; Stanford Center for Biomedical Informatics Research), Vanderbilt University (PheKB, eMERGE consortium), and the University of Washington/Fred Hutchinson Cancer Research Centre.

R13

Supplementary information

1. Describe the alignment of the project with the HTMR Network strategy

This PhD project has the potential to have a major effect on the efficiency of future trials. It seeks to tackle the methodological challenge of handling freetext adverse event information at the scale, volume and speed required for the future generation of clinical trials. This work aims to improve trial conduct, efficiency and cost-effectiveness and so fits well with the HTMR Network strategy. It will particularly focus on the development of novel, streamlined and highly cost-effective methods for use in clinical trials.

2. Does this project align with the work of a HTMR Working Group; if so, which?

This PhD aligns directly with the work of the Health Informatics group and the student will join this working group. The Trial Conduct group will also find the results of interest as the main aim of this project is to make substantial improvements in efficiency and data quality of future trials. The project has broad relevance across many areas of clinical trials work and the project will be of interest to all Hubs. It is anticipated that, via the working groups, opportunities for cross-Hub collaboration will emerge as the project proceeds.

3. Describe how this project aligns with the host Hub strategy

This fits well into the CTSU Hub strategy to achieve reliable results through supporting streamlined approaches to conducting large cost-effective randomized trials of relevance to public health. The successful candidate will be able to draw on CTSU's expertise in large-scale clinical trials, experience of using routine healthcare data (hospital admissions, death, cancer) for other trials and for UK Biobank, and integral role in data analytics capacity of the Big Data Institute. Prof Davies, who will be located in the Big Data Institute when it opens in January 2017, provides additional expertise in software engineering and the extraction, annotation and analysis of electronic healthcare records gained through his work for the 100,000 Genome Project and the NIHR Health Informatics Collaborative.

4. Detail of any Project specific training offered in the studentship

The project will provide the opportunity to be involved in collaborations across the MRC HTMR network and wider, learn to write project proposals, to attend relevant training courses and to present your research findings at relevant meetings. The project will be based in the Clinical Trial Service Unit and Epidemiological Studies Unit which has excellent facilities, and a world-class community of scientists and statisticians.

5. Are there any prerequisite qualifications or experience for this studentship?

Candidates for an MRC-funded studentship must meet residence eligibility and hold qualifications in a relevant subject at the level of, or equivalent to, a good honours degree from a UK academic institution (see methodology website for more details- www.methodologyhubs.mrc.ac.uk).

For this project: The successful applicant is expected to have a good grounding in computer science, and will work closely with experienced clinicians, statisticians, data analysts and software engineers.

A PhD candidate at the Nuffield Department of Population Health needs to demonstrate:
1. Proven academic excellence (i.e., 1st class or upper second-class undergraduate degree; or international equivalent)
2. Proficiency in English and excellent communications skills
3. Research or employment experience relevant to population health would be beneficial.