Statistical challenges in utilising Electronic Health Records for medical research

Wednesday 26th September 2018

This meeting will focus on statistical challenges in the use of Electronic Health Records for prognostic modelling and observational epidemiology. The meeting will showcase some of the wide ranging statistical methods that are being developed and applied to large-scale routinely collected clinical data

Speakers:

Elizabeth Williamson (London School of Hygiene and Tropical Medicine)

John Tazare (London School of Hygiene and Tropical Medicine)

Harvey Goldstein (University of Bristol)

Angela Wood (University of Cambridge)

Peter Diggle (Lancaster University and Health Data Research UK)

Venue: The Hardy Room, De Morgan House, 57-58 Russell Square, London WC1B 4HS

Date: 1.30-5.00pm, Wednesday 26th September 2018

Registration fees: £25 for full or retired members; £10 for student members; £50 for non-members or £70 for non-members including Biometric Society membership for 2018

Organiser: Michael Sweeting (contact via the webpage https://biometricsociety.org.uk/committee)

Programme

-

13:30 - 14:05	**Elizabeth Williamson (LSHTM)** ***Missing data in electronic health record research: challenges and (some) solutions*** Studies conducted using routinely collected data - such as electronic health records (EHR) - often suffer from missing data. While the field of missing data methodology is comparatively well developed, certain features of EHR data mean that the application of missing data methods requires specific consideration in this setting. One key issue is the sheer size of the available data, with large numbers of patients, but also large amounts of data collected on each patient over many years. A second issue is that data were not collected for the purpose of research, leading to data-dependent sampling: measurements are often only taken when the patient visits their GP, an action which is often prompted by reasons related to the health status of the patient at that time. This talk will explore problems of missing data in studies using EHR data, paying particular attention to features specific to this setting. A range of possible missing data methods will be discussed, and illustrated on a range of real EHR studies.
14:05 - 14:40	**John Tazare (LSHTM)** ***Application of the high-dimensional Propensity Score (hd-PS) to UK EHRs*** The hd-PS algorithm aims to account for confounding by adjusting for a large set of proxies thought to be informative of the underlying health status of patients. Evidence in US claims data (where the algorithm was developed) suggesting the algorithm outperforms investigator lead propensity score approaches and has led to its widespread adoption in a variety of contexts. However, the potential disparity between claims databases and UK EHRs means careful consideration of how source-specific characteristics are handled is necessary to avoid implementing the algorithm in a manner discordant to hd-PS principles. Using a recent study from the Clinical Practice Research Datalink we will present results incorporating a series of modifications aiming to tailor hd-PS principles to this setting.
14:40 - 15:15	**Harvey Goldstein (University of Bristol)** ***Linkage matching errors and quantifying uncertainty for model fitting*** The talk will describe how the inherent uncertainty about the correctness of record matches can be quantified and utilised in subsequent statistical modelling. The proposal is to extend a Bayesian missing data algorithm to incorporate data priors estimated from a particular record matching algorithm. Implications of this approach for the development of population record 'spines' will be mentioned.
15:15 - 15:45	**Tea/Coffee Break**

15:45 - 16:20	**Angela Wood (University of Cambridge)** ***Estimating cardiovascular disease risk in electronic health records with incomplete records and repeated measurements of risk predictors*** Stratification of individuals according to their estimated cardiovascular disease (CVD) risk is used to guide clinical decision-making. Current UK guidelines for CVD risk assessment recommend the use of already recorded risk factors in electronic primary care records to prioritise patients for a full formal risk assessment, although there is no guidance on how this should be achieved. We present a computationally feasible statistical approach to address the methodological challenges in utilizing historical repeat risk factors measures recorded in primary care records to systematically identify patients at high risk of future CVD disease. The approach is principally based on a dynamic two-stage landmark model. The first stage estimates predicted current risk factor values from all available historical repeat risk factor measurements by landmark-age-specific multivariate linear mixed-effects models with correlated random-intercepts, which account for sporadically recorded repeat measures, unobserved data and measurements errors. The second stage predicts future disease risk from a sex-stratified Cox proportional hazards model, with predicted current risk factor values estimated from the first stage. We have developed and internally validated a dynamic 10-year cardiovascular disease risk prediction model using electronic primary care records for age, diabetes status, hypertension treatment, smoking status, systolic blood pressure, total and high-density lipoprotein cholesterol from ~2 million individuals in ~400 primary care practices in England and Wales contributing to Clinical Practice Research Datalink. We propose using such models as pre-screening tools for identifying individuals who may be at greatest health need of a more formal CVD assessment. Using public health modelling, we identify optimal pre-screening risk thresholds for inviting individuals in for a formal risk assessment.

16:20 - 16:55

Peter Diggle (University of Lancaster and Health Data Research UK)
Real-time spatial health surveillance using routinely recorded clinical data

Spatially and temporally referenced data on a wide range of health outcomes are routinely recorded within NHS IT systems. These data present obvious opportunities for monitoring geographical variations in incidence to establish the expected pattern of geographical variation and to detect departures from expectation in real-time.

A natural modelling framework for data of this kind is a spatio-temporal point process. Ideally, this requires data to be available on all, or at least a completely random sub-set of, incident cases with finely resolved spatial and temporal coordinates of each case. In practice, spatial resolution is often constrained by concerns about confidentiality of individual-level information, or data are available only from a “sentinel network” such as a sample of general practices within the geographical region of interest.