Three Controversies in Data Science for Medicine and Healthcare

Dr Niels Peek  
University of Manchester  
Farr Institute of Health Informatics Research  
United Kingdom

Dr Pedro Pereira Rodrigues  
Center for Health Technology and Services Research  
University of Porto  
Portugal
Health Informatics: building bridges
Voting website

sli.do + #2974
Data shall be used only for the purpose for which they were collected

Big Data and predictive analytics should replace randomised clinical trials

To accelerate research, all medical and healthcare data should made available to data scientists
Controversy 1

Data shall be used only for the purpose for which they were collected

(J. van der Lei, 1991)

Please enter your vote at

sli.do + #2974
Electronic health records

- **Why health records?**
  - direct care
  - reimbursement
  - legal obligation
  - protection against lawsuits

- **Increasingly kept electronically**
  - e.g. U.S. HI-TECH act

- This offers **unprecedented opportunities** use these data for research
Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians

- The Australian Medicare system has records of health services for all Australians
- Electronic Medicare records were accessed to identify all Australians aged 0-19 years on 1-1-1985, or born between 1-1-1985 and 31-12-2005
- The cohort was followed to 31 December 2007 by linkage to the Australian Cancer Database and the National Death Index
Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians

- The mean duration of follow-up after exposure was 9.5 years.
- Overall cancer incidence was 24% greater for people exposed to CT scanning (P<0.001)
- (Corrected for age, sex, and year of birth.)
Health state information derived from secondary databases is affected by multiple sources of bias

Darcey D. Terris a,b,*, David G. Litaker a,c, Siran M. Koroukian a

aDivision of Health Services Research & Policy, Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA
bDepartment of Tropical Hygiene and Public Health, School of Medicine, University of Heidelberg, Heidelberg, Germany
cCenter for Quality Improvement Research, Louis Stokes Cleveland Department of Veterans Affairs Medical Center, Cleveland, Ohio, USA

Accepted 8 August 2006

Abstract

Objective: Secondary databases are used in descriptive studies of patient subgroups; evaluation of associations between individual characteristics and diagnosis, prognosis, and/or service utilization rates; and studies of the quality of health care delivered. This article identifies sources of bias for health state characteristics stored in secondary databases that arise from patients' encounters with health systems, highlighting sources of bias that arise from organizational and environmental factors.

Study Design and Setting: Potential sources of bias, from patient access to services and diagnosis, through encoding and filing of patient information in secondary databases, are discussed. A patient presenting with acute myocardial infarction is used as an illustrative example.

Results: The accuracy of health state characteristics derived from secondary databases is a function of both the quality and quantity of information collected before data entry and is dependent on complex interactions between patients, clinicians, and the structures and systems surrounding them.

Conclusion: The use of health state information included in secondary databases requires that estimates of potential bias from all sources be included in the analysis and presentation of results. By making this common practice in the field, greater value can be achieved from secondary database analyses. © 2007 Elsevier Inc. All rights reserved.

Keywords: Administrative data; Databases; Medical record system; Health status; Methodology; Risk adjustment
Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research

Alexander Rusanov1†, Nicole G Weiskopf2†, Shuang Wang3 and Chunhua Weng2*

Abstract

**Background:** To demonstrate that subject selection based on sufficient laboratory results and medication orders in electronic health records can be biased towards sick patients.

**Methods:** Using electronic health record data from 10,000 patients who received anesthetic services at a major metropolitan tertiary care academic medical center, an affiliated hospital for women and children, and an affiliated urban primary care hospital, the correlation between patient health status and counts of days with laboratory results or medication orders, as indicated by the American Society of Anesthesiologists Physical Status Classification (ASA Class), was assessed with a Negative Binomial Regression model.

**Results:** Higher ASA Class was associated with more points of data: compared to ASA Class 1 patients, ASA Class 4 patients had 5.05 times the number of days with laboratory results and 6.85 times the number of days with medication orders, controlling for age, sex, emergency status, admission type, primary diagnosis, and procedure.

**Conclusions:** Imposing data sufficiency requirements for subject selection allows researchers to minimize missing data when reusing electronic health records for research, but introduces a bias towards the selection of sicker patients. We demonstrated the relationship between patient health and quantity of data, which may result in a systematic bias towards the selection of sicker patients for research studies and limit the external validity of research conducted using electronic health record data. Additionally, we discovered other variables (i.e., admission status, age, emergency classification, procedure, and diagnosis) that independently affect data sufficiency.
How is the electronic health record being used? Use of EHR data to assess physician-level variability in technology use

Jessica S Ancker,1,2 Lisa M Kern,1,2 Alison Edwards,1,2 Sarah Nosal,3 Daniel M Stein,1 Diane Hauser,3 Rainu Kaushal,1,2 with the HITEC Investigators

ABSTRACT

Background Studies of the effects of electronic health records (EHRs) have had mixed findings, which may be attributable to unmeasured confounders such as individual variability in use of EHR features.

Objective To capture physician-level variations in use of EHR features, associations with other predictors, and usage intensity over time.

Methods Retrospective cohort study of primary care providers eligible for meaningful use at a network of federally qualified health centers, using commercial EHR data from January 2010 through June 2013, a period during which the organization was preparing for and in the early stages of meaningful use.

Results Data were analyzed for 112 physicians and nurse practitioners, consisting of 430,803 encounters.

Provider-level variability was high: for example, the annual average proportion of encounters with problem lists updated ranged from 5% to 60% per provider. Some metrics were associated with provider, patient, or encounter characteristics. For example, problem list updates were more likely for new patients than established ones, and alert acceptance was negatively correlated with alert frequency.

Conclusions Providers using the same EHR developed personalized patterns of use of EHR features. We conclude that physician-level usage of EHR features may be a valuable additional predictor in research on the effects of EHRs on healthcare quality and costs.
Example: prevalence of diabetes in the UK
Summary: Why is it a good idea to re-use EHR data for research?

- Allows us to answer research questions that otherwise would remain unanswered
- We can achieve much larger numbers than with conventional studies
- ... against much lower cost
Summary: Why is it not a good idea to re-use EHR data for research?

- Many sources of bias and uncertainty
- Huge variation in data quality
- Few methods to assess data quality
Controversy 1

Data shall be used only for the purpose for which they were collected

(J. van der Lei, 1991)

Please enter your vote at
sli.do + #2974
Data shall be used only for the purpose for which they were collected – posterior poll

- Agree
- Do not agree

Voting closed

Switch event
Controversy 2

Big Data and predictive analytics should replace randomised clinical trials

Please enter your vote at
sli.do + #4169
Randomised clinical trials

1937  Elixir Sulfanilamide kills >100 patients
1938  US FDC Act mandates pre-market safety evaluation
1948  First published RCT ("Streptomycin in treatment of pulmonary tuberculosis")
1961  Thalidomide causes severe birth defects and deaths in thousands
1962  Legislation mandates FDA approval contingent on “substantial evidence” of safety (first in animals and then humans) in addition to efficacy
Clinical trial costs

Figure 3: Clinical Trial Costs (in $ Millions) by Phase and Therapeutic Area

Source: US Department of Health and Human Services, 2014
Number of trials grows exponentially

Limitations of RCTs

- There are many situations where we cannot use RCTs (e.g. blinding not possible)
- RCTs are inappropriate to detect long-term effects and rare side-effects
- RCTs are not representative of clinical practice

Box 1. Hierarchy of Study Designs for Intended Effects of Therapy
1. Randomised controlled trials
2. Prospective follow-up studies
3. Retrospective follow-up studies
4. Case-control studies
5. Anecdotal: case report and series
The big target here isn't advertising, though. It's science. The scientific method is built around testable hypotheses. [...] Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). [...] But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.
The system would query a universal, de-identified clinical database in real time; identify prior cases of sufficient similarity; and provide decision support such as suggested interventions, based on prior outcomes.
“Machine learning does not solve any of the fundamental problems of causal inference in observational data sets. Algorithms may be good at predicting outcomes, but predictors are not causes. The usual commonsense caveats about confusing correlation with causation apply; indeed, they become even more important as researchers begin including millions of variables in statistical models.”
Causation vs association

Hemán, J Epidemiol Community Health 2004;58:265–271.
Confounding

A lack of comparability between exposed and unexposed groups arising because, had the exposed actually been unexposed, their disease risk would have been different from that in the actual unexposed group.

Confounders

- be a cause of the disease, or a surrogate measure of a cause, in unexposed people
- be correlated with exposure in the study population
- not be an intermediate step in the causal pathway between the exposure and the disease
Summary: Why Big Data and analytics should replace RCTs

- Randomised clinical trials are too expensive
- They do not generalise well to the real world
- There are many situations where we cannot use RCTs
- Pragmatic, data-driven approaches work better (when there is sufficient data)
Summary: Why Big Data and analytics should **not** replace RCTs

- There is no substitute for randomisation when it comes to causal inference.
- The reason is that we do not know which confounders we do not know.
- Even Google understands that very well: they A/B-test everything.
- Essential when it concerns medical interventions.
Controversy 2

Big Data and predictive analytics should replace randomised clinical trials

Please enter your vote at sli.do + #4169
Data shall be used only for the purpose for which they were collected – posterior poll

- Agree
- Do not agree

Voting closed

Switch event
Controversy 3

To accelerate research, all medical and healthcare data should made available to data scientists

Please enter your vote at sli.do + #9715
The system would query a **universal, de-identified clinical database** in real time; identify prior cases of sufficient similarity; and provide decision support such as suggested interventions, based on prior outcomes.
A movement for cancer patients, which aims to explain the risks but also to build confidence in the benefits of using of patient data for analysis and research.
Public preferences for electronic health data storage, access, and sharing — evidence from a pan-European survey

Sunil Patil,¹ Hui Lu,¹ Catherine L Saunders,¹ Dimitris Potoglou,² and Neil Robinson¹

ABSTRACT

Objective To assess the public’s preferences regarding potential privacy threats from devices or services storing health-related personal data.

Materials and Methods A pan-European survey based on a stated-preference experiment for assessing preferences for electronic health data storage, access, and sharing.

Results We obtained 20,882 survey responses (94,606 preferences) from 27 EU member countries. Respondents recognized the benefits of storing electronic health information, with 75.5%, 63.9%, and 58.9% agreeing that storage was important for improving treatment quality, preventing epidemics, and reducing delays, respectively. Concerns about different levels of access by third parties were expressed by 48.9% to 68.8% of respondents.

- 20,882 survey responses from 27 EU member countries
- Respondents were strongly averse to health insurance companies, pharma companies, and academic researchers viewing their data
care.data

- Controversial initiative to create single database with all English primary care EHR data (64.1M population)
- Legal basis provided by Health and Social Care Act 2012
- Major public concerns about informed consent and the default ‘opt-in’; privacy and data security; and involvement of private companies
- Closed since July 2016
Summary: Why all data should be made available to data scientists

- Massive increase in efficiency of health research
- No alternative to study rare diseases and rare events (CT scans example)
- Patients ask for it
- There exist methods to share data while maintaining privacy
- Public opinion about privacy is shifting anyway
Summary: Why all data should not be made available to data scientists

- Surveys indicate that citizens have major concerns about health data sharing.
- Disclosure of health data can have major societal implications for individuals (e.g. sexually transmittable diseases, mental illness).
- Risk of increasing health inequalities if data becomes available to insurance companies.
- Would undermine trust of the general public in Data Science.
Controversy 3

To accelerate research, all medical and healthcare data should be made available to data scientists.

Please enter your vote at slido + #9715
Thank you for listening!

Was this useful?

Please enter your vote at sli.do + #9715

Dr Niels Peek  
niels.peek@manchester.ac.uk

Dr Pedro Pereira Rodrigues  
pprodrigues@med.up.pt
Informatics for Health 2017
Joint meeting of MIE 2017 and The Farr Institute International Conference 2017

Venue: Manchester, UK
Date: 24th - 26th April 2017
Web: www.informaticsforhealth.org

#lforH2017