Methods for enhancing the reproducibility of observational research using electronic health records: preliminary findings from the CALIBER resource

Spiros Denaxas, Arturo Gonzalez-Izquierdo, Maria Pikoula, Kenan Direk, Natalie Fitzpatrick, Harry Hemingway
University College London, London, UK
{s.denaxas, m.pikoula, k.direk, n.fitzpatrick, h.hemingway, a.gonzalez-izquierdo}@ucl.ac.uk

Liam Smeeth
London School of Hygiene and Tropical Medicine
Electronic Health Records Research Group
London, UK
Liam.Smeeth@lshtm.ac.uk

***Abstract*—The ability of external investigators to reproduce published scientific findings is critical for the evaluation and validation of health research by the wider community. However, a substantial proportion of health research using electronic health records, data collected and generated during routine clinical care, potentially cannot reproduced. With the complexity, volume and variety of electronic health records made available for research steadily increasing, it is critical to ensure that findings from such data are reproducible and replicable by researchers. In this paper, we present some preliminary findings on how a series of methods and tools utilized in adjunct scientific disciplines can be used to enhance the reproducibility of research using electronic health records.**

Keywords; electronic health records, reproducibility

I. INTRODUCTION

The replication of studies by independent investigators, methods and datasets is one of the cornerstones of how scientific findings are evaluated and validated by the wider scientific community. While overall preclinical research probably has a much larger challenge of nonreproducibility, a significant proportion of clinical research using routinely collected health data, e.g. electronic health records (EHR), administrative health data, and disease registries, cannot be replicated [1].

Nonreproducibility can occur for several reasons but perhaps the main one is the fact that EHR are collected for patient care and not research. The quality and complexity of EHR data greatly varies across healthcare settings and disease areas. As a result, EHR data require a substantial amount of preprocessing in order to be transformed into research-ready datasets that can be statistically analyzed [2]. These data transformation operations however are not performed in a systematic manner and details are rarely included in published scientific literature. Critical information, such as for example, implementation details of EHR-derived phenotyping algorithms are not routinely nor systematically made available. As a result, reproducing scientific findings from such data is challenging. Progress however has been achieved by the creation of reporting guidelines, such as the Reporting of Studies Conducted Using Observational Routinely-collected Data (RECORD) [3] which contain a checklist of reporting items for research conducted using EHR.

In this paper, we present preliminary results of how methods/tools used in adjunct scientific domains can enable researchers to actuate the principles behind RECORD and describe how these have been used in CALIBER to facilitate the internal and external reproducibility of findings.

II. METHODS

A. CALIBER

CALIBER [4] is a research platform, established in 2009, linking national, structured primary care, hospital care, disease registry, mortality EHR data in the UK for 10m patients. Primary care data is provided by the Clinical Practice Research Datalink, an anonymized national cohort of longitudinal data for all individuals registered with a General Practitioner (GP) and recorded using the Read controlled clinical terminology (a subset of SNOMED-CT). Secondary care data is obtained from Hospital Episode Statistics (HES), a national database of administrative data used for hospital reimbursement and recorded in ICD-10. Finally, mortality and socioeconomic status data are curated by the Office of National Statistics (ONS)

and recorded in ICD-9 and ICD-10. One of the main aims of CALIBER is to enable the reproducible and transparent use of EHR for research.

B. Methods review and evaluation

We systematically searched literature and internet resources for established approaches used in computer science, biomedical informatics, bioinformatics, computational biology, and software engineering. We evaluated the identified solutions against the relevant reporting items in RECORD 1) The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail (RECORD 6.1); 2) A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided (RECORD 7.1); 3) Describe in detail the selection of the persons included in the study (i.e., study population selection) including filtering based on data quality, data availability and linkage (RECORD 13.1)

III. Results

We identified a series of tools and approaches that can potentially be utilized to enable reproducible research using EHR and describe preliminary findings below.

A. Version control systems

Research using EHR invariably generates a substantial amount of programming code across pre-processing and statistical analysis stages. In CALIBER, the majority of data manipulation operations and disease phenotyping algorithms are performed within a relational database management system and are expressed using SQL while statistical analysis is undertaken using applications such as R (https://cran.r-project.org/). Even the slightest changes, accidental or intended (e.g. updating an exposure definition), in the code can have large consequences in findings. Given the collaborative, incremental and iterative nature of research using EHR [5], it is essential for researchers to have the ability to track changes in disease or study population definitions over time and share the code used.

The standard solution for tracking the evolution of code over time is to use a version control system such as Git (https://git-scm.com/). Version control systems, widely used in software engineering, are applications that enable the structured tracking of changes to individual text-based files both over time and across multiple users. In CALIBER, the SQL code for generating study populations and phenotyping algorithms for each study is stored within a private version control system. This enables researchers to keep track of changes in study definitions at the desired time granularity and facilitates the collaborative creation of algorithms.

B. Literate programming

A typical research project involves a number of statistical analyses being performed and their output interpreted. While both of these operations are highly interconnected logically, the textual representation of the results and the code used to statistically analyze the data and produce the results are not. One way we have approached this challenge and integrate analytical code and textual context with results and their interpretation is the use of literate programming techniques [6]. Literate programming is a programming paradigm introduced by D. Knuth where compilable computer source code is interspersed with narrative in natural language. This enables researchers to provide the contextual information on a particular decision (e.g. to exclude patients with a particular set of diagnostic codes from their study population) with the actual code that executes that. Multiple literate programming packages exist for common programming languages used such as knitr (https://yihui.name/knitr/) for R and Jupyter Notebook (https://ipython.org/notebook.html) for Python.

C. Standardized analytical methods

Statistical methods are often not adequately documented or shared and are not standardized. Common data manipulations on EHR are repeated by researchers but neither code nor data are systematically shared. In contrast with genomics, where related technologies have been deemed essential, such approaches are not been widely

adopted. For example, Bioconductor (https://www.bioconductor.org/) is an open-source, open-development software project for the analysis and comprehension of high-throughput data in genomics and molecular biology based on R and contains >900 packages. Similar approaches in EHR research, such as rEHR [7] can enable the standardization of algorithms and analytical approaches and their transparent sharing as part of the dissemination process.

D. Others methods and approaches

We additionally identified other complementary methods and approaches to ones described above such as the use of test-driven development techniques, metadata standards for curating data and analytical methods and modular software engineering approaches which are not discussed in this manuscript.

IV. CONCLUSION

The challenge of reproducibility has been widely recognized by the scientific community [8]. In this paper we presented some preliminary findings of how approaches and methods used in adjusted scientific disciplines can be utilized to enable reproducible and transparent research using EHR and enable researchers to actuate reporting guidelines. Enabling reproducible research using EHR is an ongoing process but will benefit the scientific and wider community.

REFERENCES

[1] J.P.A Ioannides, "Acknoweldging and overcoming nonreproducibility in basic and preclinical research", JAMA, 2017, doi:0.1001/jama.2017.0549.

[2] S. Denaxas and K. Morley, "Big biomedical data and cardiovascular disease research: opportunities and challenges", European Heart Journal: Quality of Care and Clinical Outcomes, vol. 1, 2015, pp. 9-16, doi: 10.1093/ehjqcco/qcv005.

[3] E. Benchimol, L. Smeeth, A. Guttman, K. Harron, D. Moher, I. Petersen, H. Sorensen, E. von Elm, S. Langan, "The Reporting of Studies Conducted using Observational Routinely-collected health Data (RECORD) Statement", PLOS Med., 2015, doi: 10.1371/journal.pmed.1001885

[4] S. Denaxas, J. George, E. Herrett, A. Shah, D. Kalra, A. Hingorani, M. Kivimaki, A. Timmis, L. Smeeth, H. Hemingway, "Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)", Int J Epidemiol, vol. 41, 2012, pp. 1625-38, doi: 10.1093/ije/dys188.

[5] S. Denaxas, J. Wallace, S. Denaxas, R. Hunter, R. Patel, P. Perel, A. Shah, A. Timmis, R. Schilling, H. Hemingway, "Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation", PLOS ONE, 2014, doi: 10.1371/journal.pone.0110900.

[6] D. Knuth, "Literate programming", CSLI Lecture Notes, 1992.

[7] D. Springate, R. Parisi, I. Olier, D. Reeves, E. Kontopantelis, "rEHR: An R package for manipulating and analysing EHR data", PLOS ONE, 2017, doi: doi:10.1371/journal.pone.0171784

[8] V. Stodden, P. Guo, Z. Ma, "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals", PLOS ONE, 2013, doi: 10.1371/journal.pone.0067111.