Title: Transparency in the use of coded healthcare data for published studies in the context of heart failure

Review Question

Electronic healthcare records (EHR) and other coded healthcare data are increasingly used to determine disease status in epidemiological studies, clinical trials, and for healthcare quality assessment and improvement. However, there is a lack of transparency about how medical conditions, including underlying diseases, comorbidities and outcomes are defined in such studies, undermining the value of the resulting scientific findings and limiting the possibility of external validation.[1] This study aims to describe the trends in utilisation of coded health record data over a six-year period using heart failure (HF) as an exemplar, describing whether studies have openly disclosed their use of coded healthcare data. The process will be conducted using manual curation of 200 papers to train a machine learning approach (natural language processing [NLP]), upscaling to cover a large volume of articles across a wide breadth of journals.

Our objectives are to:

1. Evaluate the use of coded healthcare data across human subject research assessing HF.
2. Calculate the proportion of studies that are transparent about their use of coded healthcare data (e.g. adequate reporting of the use of EHR, medical claims or registry data).
3. Assess transparency in the reporting of how coded healthcare data were used, including dataset construction, linkage and coding schemes.
4. Compare the terminology of coding used to define HF, and where available, the coding schemes and code lists used.

Searches

EMBASE and MEDLINE databases will be searched from 1st January 2015 to 31st December 2020. A broad search description for HF will be used to identify relevant studies:

'acute heart failure'/exp/mj OR 'congestive heart failure'/exp/mj OR 'heart ventricle failure'/exp/mj OR 'cardiopulmonary insufficiency'/exp/mj OR 'systolic dysfunction'/exp/mj OR 'diastolic dysfunction'/exp/mj

OR

'heart failure':ab,ti OR 'heart ventricular failure':ab,ti OR 'cardiopulmonary insufficiency':ab,ti OR 'systolic dysfunction':ab,ti OR 'diastolic dysfunction':ab,ti

Types of study to be included

Following exclusion of journals focused on reviews, case reports, intensive care, basic research, paediatric care and imaging, studies will be included from journals meeting the following criteria:

(1) Availability to extract and share XML data for NLP purposes;

(2) Within the top 25 available journals, based on their impact factor rating (Clarivate Analytics 2019 categories: 'Cardiac & Cardiovascular Systems' and 'Medicine, General & Internal'): Journal of the American Medical Association (JAMA), European Heart Journal, JAMA Internal Medicine, Diabetes Care, JAMA Cardiology, European Journal of Heart Failure, Cardiovascular Diabetology, Clinical Journal of the American Society of Nephrology, European Journal of Preventive Cardiology, Clinical Research in Cardiology, Heart, Open Heart, JAMA Network Open, Journal of the American Heart Association, Journal of Hypertension, Cardiovascular Drugs and Therapy, Europace, ESC Heart Failure, European Heart Journal Acute Cardiovascular Care, European Journal of Clinical Investigation, Journal of Cardiovascular Translational Research, PLOS One, Disease Markers, Journal of Clinical Hypertension, American Journal of Hypertension.

Animal studies and studies not available in English will be excluded.

Condition or domain being studied

Published studies using coded healthcare data in human participants.

Participants/population

Focus on studies related to patients with HF (study defined), or describing HF as a comorbidity or outcome.

Intervention/exposure

Use of coded healthcare data, for example ICD, SNOMED or READ codes, or any other coding scheme for clinically-acquired healthcare data.

Comparator/control

Not applicable.

Main outcomes

Proportion of studies of human subject studies using coded healthcare data to define disease or ascertain outcomes.

Proportion of studies with clear unambiguous statements about their use of coded healthcare data to define disease or ascertain outcomes.

Measures of effect

Summary and descriptive statistics.

Additional outcomes

1. Summary and descriptive comparison of study origin, design, sample size and type of HF.
2. Proportion of coded healthcare studies providing a clear description of dataset construction and data linkage.
3. Proportion of structured healthcare data studies providing coding schemes and lists used to define HF.
4. Descriptive comparison of coding schemes used to define HF.

Data extraction (selection and coding)

200 random papers from the search list will each be assessed by 2 reviewers independently, with consensus discussion to resolve discrepancies, and if necessary third person adjudication. These findings will be used to train the NLP for automated

extraction in the main body of the paper in up to 5000 journal articles. A further 420 random papers will be used to train and validate the NLP model.

Risk of bias (quality) assessment

A proprietary system will be employed to evaluate the quality of the 200 random studies in the manual curation process. This is derived from the global multi-stakeholder CODE-EHR framework for the use of structured healthcare data in research.[2] These items also match the RECORD checklist items in section 6.1, 6.2 and 7.1.

1. Published protocol available (yes = low risk of bias; no = high risk of bias).
2. Clear description of dataset construction and any linkage performed (yes = low risk of bias; no = high risk of bias).
3. Sufficient detail on how diseases, conditions and outcomes were defined, including those relating to patient identification, therapy, procedures, comorbidities or adverse events (yes = low risk of bias; no = high risk of bias).
4. Sufficient detail on validation and the analytical processes undertaken, including use of algorithms and machine learning approaches (yes = low risk of bias; no = high risk of bias).
5. Ethical governance, with clear unambiguous statements on how the principles of Good Clinical Practice and Data Protection were met (yes = low risk of bias; no = high risk of bias).

Strategies for data synthesis/analysis

A standardised data extraction form will be used for the manual curation process. The extraction from each pair of reviewers will be amalgamated after consensus is reached on any discrepancies. Summary and descriptive statistics will be used to evaluate studies where coded healthcare data were used.

The Komenti semantic text mining framework will be used for NLP[3-5]. A classifier will be built to predict the relevant data points using the stipulated terms and associated context. The predictions will be evaluated in the context of a binary classifier, producing precision, recall, and F1 values using a gold-standard subset, and contrasted using inter-annotator agreement. Ground truth for the NLP analysis will be derived from the manual analysis of a subset of the literature documents, and therefore the classifier outcomes will be defined by the rubric used for manual extraction. A keyword approach will be used to match relevant phrases in the document, which will then be synthesised into a binary outcome using multiple context disambiguation methods (such as negation and uncertainty detection). Subsequently, NLP will be explored for identification of additional outcomes recorded by the manual extraction process, such as geographic area, type of HF, and others. This portion of the study will be exploratory and will consist of further deployment of the keyword and context disambiguation approach, as well as novel research into vectorisation and similarity-based approaches in a supervised manner, for identification of more complex outcomes, using a portion of the manually extracted data as training. This could also extend into detection of good or bad practices for reporting of coded healthcare data use, using output from the manual curation of examined documents.

Where data are sufficient, analysis will be stratified according to the stated type of HF (reduced vs preserved ejection fraction), geographic location of the study (Europe; USA; Latin America/Canada; Asia-Pacific, Middle East and Africa; or multiple regions) and coding system used (ICD, SNOMED, etc).

Type and method of review

Systematic review

Health area of review

Cardiovascular

Country

Global

Review team members and their organisation affiliations

List all team members and affiliated institutions

Dr Luke Slater, University of Birmingham

Dr Asgher Champsi, University of Birmingham

Dr Simrat Gill, University of Birmingham

Dr. Tomasz Dyszynski, Bayer AG

Megan Molar, Bayer AG

Dr. Kiliana Suzart-Woischnik, Bayer AG

Dr. Benoit Tyl, Bayer AG

Dr. Guillaume Allee, Servier

Dr. Alfonso Sartorius, Servier

Dr Tom Lumbers, University College London

Professor Georgios Gkoutos, Universty of Birmingham & University Hospitals Birmingham NHS Trust

Professor Dipak Kotecha, University of Birmingham & University Hospitals Birmingham NHS Trust

Anticipated or actual start date
October 2022

Anticipated completion date
August 2023

References

1. R Studer, C Sartini, K Suzart-Woischnik, R Agrawal, H Natani, SK Gill, SB Wirta, FW Asselbergs, R Dobson, S Denaxas, D Kotecha. Identification and Mapping Real-World Data Sources for Heart Failure, Acute Coronary Syndrome, and Atrial Fibrillation. *Cardiology*. 2022;147:98-106. https://doi.org/10.1159/000520674.

2. D Kotecha, FW Asselbergs, European Society of Cardiology, BigData@Heart consortium, CODE-EHR international consensus group. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. 2022: Published in [1] *BMJ 2022;378:e069048 https://doi.org/10.1136/bmj-2021-069048*; [2] *Lancet Digit Health* 2022;4:e757-e764 https://doi.org/10.1016/S2589-7500(22)00151-0; and [3] *Eur Heart J* 2022;43:3578-3588 https://doi.org/10.1093/eurheartj/ehac426.

3. LT Slater, W Bradlow, T Desai, A Aziz, F Evison, S Ball, GV Gkoutos. Making Words Count with Computerised Identification of Hypertrophic Cardiomyopathy Patients. *medRxiv*. 2021:2021.04.13.21255353. https://doi.org/10.1101/2021.04.13.21255353.

4. LT Slater, W Bradlow, R Hoehndorf, DF Motti, S Ball, GV Gkoutos. Komenti: A semantic text mining framework. *bioRxiv*. 2020:2020.08.04.233049. https://doi.org/10.1101/2020.08.04.233049.

5. LT Slater, W Bradlow, DF Motti, R Hoehndorf, S Ball, GV Gkoutos. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. *Comput Biol Med*. 2021;130:104216. https://doi.org/10.1016/j.compbiomed.2021.104216.