Phenotype algorithm and data source reporting in top clinical journals: where we are and where should we go?

Anna Ostropolets, MD[1], RuiJun Chen, MD[1], Matthew Spotnitz, MD[1], Runsheng Wang, MD[1], Patrick Ryan, PhD[1,2], Prof George Hripcsak, MD[1,3]

[1]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA; [2]Janssen Research and Development, 1125 Trenton Harbourton Rd., Titusville, NJ, USA 08560; [3]New York-Presbyterian Hospital, 622 W 168 St, PH20 New York, NY 10032 USA

Research Category (please highlight or circle which category best describes your research)
Observational data management

Introduction

With a growing body of observational research there has been an increased need for transparent reporting of methods and data sources [1]. There are general standards for observational studies reporting, including RECORD-PE and STROBE frameworks [2], which provide overall guidance but do not specifically focus on the standards for phenotype or data source description. Little is known about the requirements for data provenance or phenotype reporting in existing literature.
The objective of this study was to investigate existing standards and requirements for observational data source and phenotype reporting in top clinical journals.

Methods

We searched for non COVID-19-related recent papers that use observational data in the top clinical journals (Lancet, the British Medical Journal, The Journal of the American Medical Association and JAMA Internal Medicine, New England Journal of Medicine, Circulation, Nature Medicine and Annals of Internal Medicine). From each journal we selected five papers published within the past 12 months and analyzed the description of the phenotypes and participating data sources. We focused on the description of the phenotype development process and validation, including the details about the methodology of phenotyping, their limitations and ontology use. We evaluated the description of data sources, including data provenance, elements and validation.

Results

Reporting on phenotyping development, validation and performance

More than half of the analyzed papers (23, 57.5%) relied on previously published literature in defining phenotype algorithms, even if such literature did not specify performance metrics (Table 1).

Table 1. Distribution of analyzed papers based on the type of phenotyping algorithms (original and re-used) and their validation.

	Validated phenotypes	Non-validated phenotypes	Total
Original phenotypes	2	15	17
Phenotypes constructed based on the previous studies	18	5	23
Total	20	20	40

The phenotype algorithms were mainly based on groups of ICD10 or ICD9 codes, accompanied by medication records. Other ontologies used included ATC, CPT4, Read and SNOMED. One study used unstructured data from pathology reports.
Only half of the papers used phenotype algorithms that were validated. Moreover, only 11% of original phenotypes were validated. Chart review was the main validation method with positive predictive value being the most commonly reported characteristic. Only one paper discussed misclassification error and phenotype algorithm limitations in depth.

Reporting on data source provenance and validation

Most of the papers (32, 80%) used electronic health records and administrative claims data (16 in each category), accompanied by registry data (6, 15%), hospital charge data and case report data (1, 2.5% each). Only two (5%) studies were performed on multiple data sources, which required data standardization and harmonization (Sentinel Distributed Database and OHDSI network). The data originated both from the US (18, 45%) and non-US data sources (22, 55%), mainly from Korea, Taiwan, the UK and Denmark.

The description of the data sources lacked structure and varied in the level of detail. Those studies that used well-established and previously used datasets (UK Biobank, IBM MarketScan® Medicare Supplemental Database and others) predominantly cited the previous literature describing those data sources or briefly described them. The common elements described included the type of the data (medical records or administrative claims, pharmacy records or claims, demographic data, death data), number of covered patients and time span. Fewer papers specified how the data sources were validated.
Few papers describing new data sources (for example, UK National Institute for Health Research Health Informatics Collaborative) provided a more rigorous assessment, including the institutions that contributed the data, personnel who transformed the data, details about data elements and quality assurance.
We identified the following elements that we recommend be reported for participating data sources:

- Source of the data, including participating institutions
- Data source timespan and covered population, including the number of patients, demographic and other relevant characteristics
- Details about the data elements in the data source (type of visits covered, drug prescriptions and fills, socioeconomic data and so on)
- If a data source includes multiple sources, details about their linkage and linkage evaluation
- Details of quality assurance and data validation
- Additional information about the data elements specific to the research question

An example is this description of Columbia University Irving Medical Center database:
"The Columbia University Irving Medical Center (CUIMC) database comprises electronic health records on 6,666,613 patients, with data collection starting in 1985. CUIMC is a northeast US quaternary care center with primary care practices in northern Manhattan and surrounding areas, and the database includes inpatient and outpatient care. The database currently holds information about the person (demographics), visits (inpatient and outpatient), conditions (billing diagnoses and problem lists), drugs (outpatient prescriptions and inpatient orders and administrations), devices, measurements (laboratory tests and vital signs), and other observations (symptoms). The data sources include current and previous electronic health record systems (homegrown Clinical Information System, homegrown WebCIS, Allscripts Sunrise Clinical Manager, Allscripts TouchWorks, Epic Systems), administrative systems (IBM PCS-ADS, Eagle Registration, IDX Systems, Epic Systems), and ancillary systems (homegrown LIS, Sunquest, Cerner Laboratory). The data were extracted from each system and transformed to the OHDSI OMOP Common Data Model: common data model source name "Epic Legacy CUMC MERGE," common data model ETL reference "v1.3.0.cdm5.3," common data model release date "2020-05-22," vocabulary version "v5.0 30-APR-20," with OMOP common data model version 5.3.1 and local version name "ohdsi_cumc_2020q1r4." The analysis was done 6/8/2020. A co-author has direct access to the CUIMC OMOP database."

Such a narrative can be accompanied by the aggregated statistics gathered directly from the data source to generate a comprehensive and up-to-date description. For example, EHDEN Database Catalogue leverages Achilles profiles (https://test.ehden.eu/DatabaseDashboard) to visualize objective data source characteristics, where the amount of information supplied and level of details are fully controlled by the data owner.

Conclusion
We found that most of the phenotypes used in observational studies published by clinical journals used ICD10 or ICD9 to define patients of interest, lacked the discussion of their limitations and were not validated. Lack of validation among the studies that used original phenotypes suggests that conventional methods of phenotype validation may not be feasible or scalable. The level of detail of phenotype and data source description varied greatly and was reported inconsistently. A standardized system for developing

and validating phenotypes may improve the accuracy and consistency of observational research.

References

1. Kahn MG, Brown JS, Chun AT, et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)*. 2015;3(1):1052. Published 2015 Mar 23. doi:10.13063/2327-9214.1052
2. Langan Sinéad M, Schmidt Sigrún AJ, Wing Kevin, Ehrenstein Vera, Nicholls Stuart G, Filion Kristian Bet al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE) *BMJ* 2018; 363 :k3532