Chapter 10
Defining the Patient Cohort

Ari Moskowitz and Kenneth Chen

Learning Objectives

- Understand the process of cohort selection using large, retrospective databases.
- Learn about additional specific skills in cohort building including data visualization and natural language processing (NLP).

10.1 Introduction

A critical first step in any observational study is the selection of an appropriate patient cohort for analysis. The importance of investing considerable time and effort into selection of the study population cannot be overstated. Failure to identify areas of potential bias, confounding, and missing data up-front can lead to considerable downstream inefficiencies. Further, care must be given to selecting a population of patients tailored to the research question of interest in order to properly leverage the tremendous amount of data captured by Electronic Health Records (EHRs).

In the following chapter we will focus on selection of the study cohort. Specifically, we will review the basics of observational study design with a focus on types of data often encountered in EHRs. Commonly used instrumental variables will be highlighted—they are variables used to control for confounding and measurement error in observational studies. Further, we will discuss how to utilize a combination of data-driven techniques and clinical reasoning in cohort selection. The chapter will conclude with a continuation of the worked example started in part

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-43742-2_10) contains supplementary material, which is available to authorized users.

MIT Critical Data, *Secondary Analysis of Electronic Health Records*,
DOI 10.1007/978-3-319-43742-2_10

one of this section where we will discuss how the cohort of patients was selected for the study of arterial line placement in the intensive care unit [1].

10.2 PART 1—Theoretical Concepts

10.2.1 Exposure and Outcome of Interest

These notions are discussed in detail in Chap. 9—"Formulating the Research Question". Data mining in biomedical research utilizes a retrospective approach wherein the exposure and outcome of interest occur prior to patient selection. It is critically important to tailor the exposure of interest sought to the clinical question at hand. Selecting an overly broad exposure may allow for a large patient cohort, but at the expense of result accuracy. Similarly, being too specific in the choice of exposure may allow for accuracy but at the expense of sample size and generalizability.

The selection of an exposure of interest is the first step in determining the patient cohort. In general, the exposure of interest can be thought of as patient-centric, episode-centric, or encounter centric. This terminology was developed by the data warehousing firm Health Catalyst for their Cohort Builder tool and provides a reasonable framework for identifying an exposure of interest. Patient-centric exposures focus on traits intrinsic to a group of patients. These can include demographic traits (e.g. gender) or medical comorbidities (e.g. diabetes). In contrast, episode-centric exposures are transient conditions requiring a discrete treatment course (e.g. sepsis). Encounter-centric exposures refer to a single intervention (e.g. arterial line placement) [2]. Although encounter-specific exposures tend to be simpler to isolate, the choice of exposure should be determined by the specific hypothesis under investigation.

The outcome of interest should be identified a priori. The outcome should relate naturally to the exposure of interest and be as specific as possible to answer the clinical question at hand. Care must be taken to avoid identifying spurious correlations that have no pathophysiologic underpinnings (see for instance the examples of spurious correlations shown on http://tylervigen.com). The relationship sought must be grounded in biologic plausibility. Broad outcome measures, such as mortality and length-of-stay, may be superficially attractive but ultimately confounded by too many variables. Surrogate outcome measures (e.g. change in blood pressure, duration of mechanical ventilation) can be particularly helpful as they relate more closely to the exposure of interest and are less obscured by confounding.

As EHRs are not frequently oriented towards data mining and analysis, identifying an exposure of interest can be challenging. Structured numerical data, such as laboratory results and vital signs, are easily searchable with standard querying techniques. Leveraging unstructured data such as narrative notes and radiology reports can be more difficult and often requires the use of natural language processing (NLP) tools. In order to select a specific patient phenotype from a large, heterogeneous group of patients, it can be helpful to leverage both structured and unstructured data forms.

Once an exposure of interest is selected, the investigator must consider how to utilize one or a combination of these data types to isolate the desired study cohort for analysis. This can be done using a combination of data driven techniques and clinical reasoning as will be reviewed later in the chapter.

10.2.2 Comparison Group

In addition to isolating patients mapping to the exposure of interest, the investigator must also identify a comparison group. Ideally, this group should be comprised of patients phenotypically similar to those in the study cohort but who lack the exposure of interest. The selected comparison cohort should be at equal risk of developing the study outcome. In observational research, this can be accomplished notably via propensity score development (Chap. 23—"Propensity Score Analysis"). In general, the comparison group ought to be as large as or larger than the study cohort to maximize the power of the study. It is possible to select too many features on which to 'match' the comparison and study cohorts thereby reducing the number of patients available for the comparison cohort. Care must be taken to prevent over-matching.

In select cases, investigators can take advantage of natural experiments in which circumstances external to the EHR readily establish a study cohort and a comparison group. These so called 'instrumental variables' can include practice variations between care units, hospitals, and even geographic regions. Temporal relationships (i.e. before-and-after) relating to quality improvement initiatives or expert guideline releases can also be leveraged as instrumental variables. Investigators should be on the lookout for these highly useful tools.

10.2.3 Building the Study Cohort

Isolating specific patient phenotypes for inclusion in the study and comparison cohorts requires a combination of clinical reasoning and data-driven techniques. A close working relationship between clinicians and data scientists is an essential component of cohort selection using EHR data.

The clinician is on the frontline of medical care and has direct exposure to complex clinical scenarios that exist outside the realm of the available evidence-base. According to a 2011 Institute of Medicine Committee Report, only 10–20 % of clinical decisions are evidence based [3]. Nearly 50 % of clinical practice guidelines rely on expert opinion rather than experimental data [4]. In this 'data desert' it is the role of the clinician to identify novel research questions important for direct clinical care [5]. These questions lend themselves naturally to the isolation of an exposure of interest.

Once a clinical question and exposure of interest have been identified, the clinician and data scientist will need to set about isolating a patient cohort. Phenotype querying of structured and unstructured data can be complex and requires frequent tuning of the search criteria. Often multiple, complementary queries are required in order to isolate the specific group of interest. In addition, the research team must consider patient 'uniqueness' in that some patients have multiple ICU admissions both during a single hospitalization and over repeat hospital visits. If the same patient is included more than once in a study cohort, the assumption of independent measures is lost.

Researchers must pay attention to the necessity to exclude some patients on the grounds of their background medical history or pathological status, such as pregnancy for example. Failing to do so could introduce confounders and corrupt the causal relationship of interest.

In one example from a published MIMIC-II study, the investigators attempted to determine whether proton pump inhibitor (PPI) use was associated with hypomagnesaemia in critically-ill patients in the ICU [6]. The exposure of interest in this study was 'PPI use.' A comparison group of patients who were exposed to an alternative acid-reducing agent (histamine-2 receptor antagonists) and a comparison group not receiving any acid reducing medications were identified. The outcome of interest was a low magnesium level. In order to isolate the study cohort in this case, queries had to be developed to identify:

1. First ICU admission for each patient
2. PPI use as identified through NLP analysis of the 'Medication' section of the admission History and Physical
3. Conditions likely to influence PPI use and/or magnesium levels (e.g. diarrheal illness, end-stage renal disease)
4. Patients who were transferred from other hospitals as medications received at other hospitals could not be accounted for (patients excluded)
5. Patients who did not have a magnesium level within 36-h of ICU admission (patients excluded)
6. Patients missing comorbidity data (patients excluded)
7. Potential confounders including diuretic use

The SQL queries corresponding to this example are provided under the name "SQL_cohort_selection".

Maximizing the efficiency of data querying from EHRs is an area of active research and development. As an example, the Informatics for Integrating Biology and the Bedside (i2b2) network is an NIH funded program based at Partner's Health Center (Boston, MA) that is developing a framework for simplifying data querying and extraction from EHRs. Software tools developed by i2b2 are free to download and promise to simplify the isolation of a clinical phenotype from raw EHR data https://www.i2b2.org/about/index.html. This and similar projects should help simplify the large number of queries necessary to develop a study cohort [7].

10.2.4 Hidden Exposures

Not all exposures of interest can be identified directly from data contained within EHRs. In these circumstances, investigators need to be creative in identifying recorded data points that track closely with the exposure of interest. Clinical reasoning in these circumstances is important.

For instance, a research team using the MIMIC II database selected 'atrial fibrillation with rapid ventricular response receiving a rate control agent' as the exposure of interest. Atrial fibrillation is a common tachyarrhythmia in critically-ill populations that has been associated with worse clinical outcomes. Atrial fibrillation with rapid ventricular response is often treated with one of three rate control agents: metoprolol, diltiazem, or amiodarone. Unfortunately, 'atrial fibrillation with rapid ventricular response' is not a structured variable in the EHR system connected to the MIMIC II database. Performing an NLP search for the term 'atrial fibrillation with rapid ventricular response' in provider notes and discharge summaries is feasible however would not provide the temporal resolution needed with respect to drug administration.

To overcome this obstacle, investigators generated an algorithm to indirectly identify the 'hidden' exposure. A query was developed to isolate the first dose of an intravenous rate control agent (metoprolol, diltiazem, or amiodarone) received by a unique patient in the ICU. Next, it was determined whether the heart rate of the patient within one-hour of recorded drug administration was >110 beats per minute. Finally, an NLP algorithm was used to search the clinical chart for mention of atrial fibrillation. Those patients meeting all three conditions were included in the final study cohort. Examples of the Matlab code used to identify the cohort of interest is provided (function "Afib"), as well as Perl code for NLP (function "NLP").

10.2.5 Data Visualization

Graphic representation of alphanumeric EHR data can be particularly helpful in establishing the study cohort. Data visualization makes EHR data more accessible and allows for the rapid identification of trends otherwise difficult to identify. It also promotes more effective communication both amongst research team members and between the research team and a general audience not accustomed to 'Big Data' investigation. These principles are discussed more extensively in Chap. 15 of this textbook "Exploratory Data Analysis".

In the above mentioned project exploring the use of rate control agents for atrial fibrillation with rapid ventricular response, one outcome of interest was time until control of the rapid ventricular rate. Unfortunately, the existing literature does not provide specific guidance in this area. Using data visualization, a group consensus

was reached that rate control would be defined as a heart <110 for at least 90 % of the time over a 4-h period. Although some aspects of this definition are arbitrary, data visualization allowed for all team members to come to an agreement on what definition was the most statistically and clinically defensible.

10.2.6 Study Cohort Fidelity

Query algorithms are generally unable to boast 100 % accuracy for identifying the sought patient phenotype. False positives and false negatives are expected. In order to guarantee the fidelity of the study cohort, manually reviewing a random subset of selected patients can be helpful. Based on the size of the study cohort, 5–10 % of clinical charts should be reviewed to ensure the presence or absence of the exposure of interest. This task should be accomplished by a clinician. If resources permit, two clinician reviewers can be tasked with this role and their independent results compared using a Kappa statistic.

Ultimately, the investigators can use the 'gold standard' of manual review to establish a Receiver Operating Characteristic (ROC). An area-under the ROC curve of >0.80 indicates 'good' accuracy of the algorithm and should be used as an absolute minimum of algorithm fidelity. If the area under the ROC curve is <0.80, a combination of data visualization techniques and clinical reasoning should be used to better tune the query algorithm to the exposure of interest.

10.3 PART 2—Case Study: Cohort Selection

In the case study presented, the authors analyzed the effect of indwelling arterial catheters (IACs) in hemodynamically stable patients with respiratory failure using multivariate data. They identified the encounter-centric 'arterial catheter placement' as their exposure of interest. IACs are used extensively in the intensive care unit for beat-to-beat measuring of blood pressure and are thought to be more accurate and reliable than standard, non-invasive blood pressure monitoring. They also have the added benefit of allowing for simpler arterial blood gas collection which can reduce the need for repeated venous punctures. Given their invasive nature, however, IACs carry risks of bloodstream infection and vascular injury. The primary outcome of interest selected was 28-day mortality with secondary outcomes that included ICU and hospital length-of-stay, duration of mechanical ventilation, and mean number of blood gas measurements made.

The authors elected to focus their study on patients requiring mechanical ventilation that did not require vasopressor and were not admitted for sepsis. In patients

requiring mechanical ventilation, the dual role of IACs to allow for beat-to-beat blood pressure monitoring and to simplify arterial blood gas collection is thought to be particularly important. Patients with vasopressor requirements and/or sepsis were excluded as invasive arterial catheters are needed in this population to assist with the rapid titration of vasoactive agents. In addition, it would be difficult to identify enough patients requiring vasopressors or admitted for sepsis, who did not receive an IAC.

The authors began their cohort selection with all 24,581 patients included in the MIMIC II database. For patients with multiple ICU admissions, only the first ICU admission was used to ensure independence of measurements. The function "cohort1" contains the SQL query corresponding to this step. Next, the patients who required mechanical ventilation within the first 24-h of their ICU admission and received mechanical ventilation for at least 24-h stay were isolated (function "cohort2"). After identifying a cohort of patients requiring mechanical ventilation, the authors queried for placement of an IAC sited after initiation of mechanical ventilation (function "cohort3"). As a majority of patients in the cardiac surgery recovery unit had an IAC placed prior to ICU admission, all patients from the cardiac surgical ICU were excluded from the analysis (function "cohort4"). In order to exclude patients admitted to the ICU with sepsis, the authors utilized the Angus criteria (function "cohort5"). Finally, patients requiring vasopressors during their ICU admission were excluded (function "cohort6").

The comparison group of patients who received mechanical ventilation for at least 24-h within the first 24-h of their ICU admission but did not have an IAC placed was identified. Ultimately, there were 984 patients in the group who received an IAC and 792 patients who did not. These groups were compared using propensity matching techniques described in the Chap. 23—"Propensity Score Analysis".

Ultimately, this cohort consists of unique identifiers of patients meeting the inclusion criteria. Other researchers may be interested in accessing this particular cohort in order to replicate the study results or address a different research questions. The MIMIC website will in the future provide the possibility for investigators to share cohorts of patients, thus allowing research teams to interact and build upon other's work.

Take Home Messages

- Take time to characterize the exposure and outcomes of interest pre-hoc
- Utilize both structured and unstructured data to isolate your exposure and outcome of interest. NLP can be particularly helpful in analyzing unstructured data
- Data visualization can be very helpful in facilitating communication amongst team members

References

1. Hsu DJ, Feng M, Kothari R, Zhou H, Chen KP, Celi LA (2015) The association between indwelling arterial catheters and mortality in hemodynamically stable patients with respiratory failure: a propensity score analysis. Chest 148(6):1470–1476
2. Merkley K (2013) Defining patient populations using analytical tools: cohort builder and risk stratification. Health Catalyst, 21 Aug 2013
3. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines (2011) Clinical practice guidelines we can trust. National Academies Press (US), Washington (DC)
4. Committee on the Learning Health Care System in America and Institute of Medicine (2013) Best care at lower cost: the path to continuously learning health care in America. National Academies Press (US), Washington (DC)
5. Moskowitz A, McSparron J, Stone DJ, Celi LA (2015) Preparing a new generation of clinicians for the era of big data. Harv Med Stud Rev 2(1):24–27
6. Danziger J, William JH, Scott DJ, Lee J, Lehman L, Mark RG, Howell MD, Celi LA, Mukamal KJ (2013) Proton-pump inhibitor use is associated with low serum magnesium concentrations. Kidney Int 83(4):692–699
7. Jensen PB, Jensen LJ, Brunak S (2012) Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 13(6):395–405