Analysis of Big Healthcare Databases - Exercises

Introduction

The goal of these exercises is to explore the structure of an electronic health records (EHR)-derived data set and some of the common challenges encountered in working with healthcare-derived data. We will also practice implementing some statistical methods introduced throughout the short course to address these issues. To do this we will use a synthetic data set simulated to mimic the structure of a real EHR-derived data set. **Please note that the data we will be working with are simulated and intended for instructional purposes only.** Real EHR data generally have access restrictions due to privacy/confidentiality issues and HIPAA protections. At the end of the course I will provide links for a few public access repositories that provide real EHR data. However, these generally do require a data use agreement and thus a few steps are involved in getting access.

The synthetic data we will be working with are based on the PEDSnet study of pediatric type 2 diabetes described in class. The data are divided into four files which can be downloaded from GitHub. The four files contain data from 9,930 patients age 10-20 years who had at least one outpatient encounter between 2001 and 2019. The four files can be linked using the variable patientid.

Encounter data

This data set includes one row per outpatient encounter.

patientid: Patient ID

servicedate: Date of the encounter

age: Age in years

race: Provider-reported patient race

proc: CPT codes for procedure performed; codes 99211-99215 are for evaluation and management of established patients in an outpatient setting

diag: ICD-9 (on or before 10/31/2015) or ICD-10 (after 10/31/2015) for primary diagnosis; a few diagnosis codes of interest are T2DM ICD-9 = “250.00”; T2DM ICD-10 = “E11.9”; T1DM ICD-9 = “250.01”; T1DM ICD-10 = “E10.9”; Depression ICD-9 = “296.2”,“296.9”,“296.3”,“300.4”; Depression ICD-10 = “F32.9”,“F41.8”,“F33.9”

prov: CMS provider specialty code; a few provider codes of potential interest are General Practice = “1”, General Dermatology = “7”, Family practice = “8”, Internal Medicine = “11”, Neurology = “13”, Ophthalmology = “18”, Psychiatry = “26”, Pediatric Medicine = “37”, Endocrinology = “46”

Prescription medication data

This data set includes one row per prescription recorded on the date of an outpatient encounter included in the encounters file.

> patientid: Patient ID
>
> presdate: Date of the prescription
>
> drug: Drug class

Measures data

This data set includes one row per anthropometric or laboratory measurement recorded on the date of an outpatient encounter included in the encounters file.

> patientid: Patient ID
>
> servicedate: Date of measurement
>
> measurement: Numeric value of the measurement
>
> measuretype: Description of the laboratory or anthropometric test (height in cm, weight in kg, glucose in mg/dl, hemoglobin A1c (hba1c) in %, cholesterol (chol) in mg/dl)

Validation data

This data set includes one row per patient for 998 patients randomly selected for manual chart review to determine gold-standard type 2 diabetes status.

> patientid: Patient ID
>
> T2DMv: Type 2 diabetes status based on manual chart review (1 = T2DM, 0 = no evidence of T2DM)

Install R packages

- For these exercises you will need the *rpart*, *pROC*, *boot*, and *gee* packages.
- If you have not already, please install these packages now.

Exercises

1. **Data Quality Evaluation.** The first task in analysis of EHR data is data exploration and visualization to identify and resolve data errors. Focusing on the measures data set, we will carry out a descriptive analysis.

- Are there any observations that seem likely to be errors?
- What are some techniques we can use to identify errors?
- What are some options for handling possibly erroneous data points once they have been identified?

2. **Phenotype Extraction.** We will next explore a few alternative approaches to deriving a type 2 diabetes (T2DM) phenotype from this data set.

- Reduce the data to one observation per patient considering what data elements might be of use at the patient-level.
- Use the validation data to develop a prediction model for T2DM using logistic regression and CART.
- Apply the eMERGE T2DM rule to these data. How do the sensitivity, specificity, PPV, and NPV of these approaches compare?

3. **Missing Data.** Next we will explore missing data in an EHR-derived data set. Suppose we want to use our T2DM phenotype from exercise 2 to explore the relationship between total cholesterol and T2DM diagnosis.

- Using mean cholesterol in the one year period after baseline as our exposure measure, how much missingness is there?
- Is missingness related to any other factors in the data set?
- Use IPW with a single module or multiple modules to account for missingness in your analysis of the association between total cholesterol and T2DM.

4. **Confounding by Utilization Intensity.** Accounting for confounding due to variation in the intensity of healthcare utilization is important for EHR-based analyses.

- How much variability is there in the intensity of utilization in this data set?
- Use a measure of intensity of utilization to account for informed presence bias in an analysis of the association between depression diagnosis and T2DM.

5. **Outcome Misclassification.** For binary phenotypes we can account for error in the phenotyping process using the classic Magder and Hughes approach. We will try out implementing this method in an analysis of the association between T2DM, using the CART-derived phenotype, and depression diagnosis.

- What is the crude odds ratios for the association between T2DM (outcome) and having a depression diagnosis (exposure) with and without correction for outcome misclassification?
- Using logistic regression to adjust for demographic characteristics, what is the adjusted odds ratio for the association between T2DM (outcome) and having a depression diagnosis (exposure) with and without correction for outcome misclassification?

6. **Using Probabilistic Phenotypes.** Finally, we will explore correcting for error in our phenotype using a continuous phenotype.

- Using predicted probabilities from your logistic regression-based phenotype derived in exercise 2, estimate the association between having a depression diagnosis code and T2DM.

- How do your results change if you use the bias correction approach described in lecture vs the uncorrected results?