Chapter 7

Merging Different Datasets to Allow for a Complete Analysis (Inpatient, Outpatient, Physician Visits, Medications)

INTRODUCTION

If we want to gain a complete picture of patient treatment, we need to examine all patient encounters with the medical profession. Then, we need to divide them into episodes of care. For example, an encounter with a physician might lead to a laboratory test, which in turn could lead to inpatient (or outpatient) surgery with additional follow up visits to a physician. A chronic condition could lead to multiple patient visits to the emergency department, which could become an inpatient stay. In order to determine the total cost of such episodes, we must first be able to define them.

One of the major ways we have of defining episodes is to examine dates of care. If a physician visit occurs within a week or two of an inpatient event, it is very likely a follow up to the inpatient treatment. However, if the physician visit occurs several months after an inpatient event, it is likely for a general checkup, or it may mark the start of a new patient episode. We also want to look at the overall cost of a disease over a year’s time period given all requirements of that treatment: physician visits, medications, lab tests, and hospitalizations. The MEPS is ideal for such analyses.

In anticipation of the need to define patient episodes in subsequent chapters, in this chapter, we focus on the preprocessing necessary to combine data from different data sources. This is generally known as data fusion, where different data from different sources are combined for further investigations to extract important information. In this
chapter, we extend the discussion of the many-to-many relationships that we started in Chapter 5.

BACKGROUND

Multiple Datasets

When a study requires multiple datasets, there is, as will be shown here, considerable effort in the preprocessing. This is particularly true when multiple years of data are combined and compared, as well as multiple types of data contained within multiple datasets as is typical of data from the Medical Expenditure Panel Survey. However, most of the studies do not indicate just how the data mergers take place. (Harman, Edlund, & Fortney, 2009; Kamble & Bharmal, 2009) For example, a recent study claimed to be able to compare a patient with diabetes to an identical patient without diabetes. However, co-morbidities are more likely with diabetes, so that there is a conditional probability factor that should also be taken into consideration. Again, no real information is provided to ensure that the preprocessing is performed correctly. (Balu, 2007) In this particular study, the total expenditure is the sum of expenditures contained within each of the different datasets, and the results depend very directly on the quality of the preprocessing.

There are different types of data merges, and you must be careful to use the correct merge. There are two basic types: one-to-one merging that combines observations from two data sets into a single observation in a new data set and match merging that combines observations from two data sets into a single observation in a new data set according to the values of a variable that you can specify. This second type of merging can be complete, or it can be what is called an inner or outer join. When using this type of join, great care must be exercised, or it is possible to have too many observations (duplicates) in the merged dataset.

The one-to-one merging is the safest. However, it is not always practical given the nature of the data. We will discuss this second type of merging in detail. According to a recent study, many to many merging is still not well understood or well managed. (Asiala & Gober, 2005) Great care must be taken when merging, or the results will lead to erroneous conclusions. (Paiba, et al., 2007)

Medicare, Part D

Starting with January, 2006, Medicare, part D has funded prescription claims. Prior to that year, Medicare did not generally pay for any prescriptions. We want to investigate the impact of Medicare, part D on patient usage of prescriptions and correspondingly, patient usage of other types of health care, including inpatient, outpatient, and physician visits.

According to a recent article, the impact of Medicare, Part D remains speculative. (Hoffman, et al., 2009) There is some indication that more seniors without previous prescription coverage are increasing their utilization of medications. (Hudson, 2009) What the few studies that are available based upon actual data have tended to be global, looking only at total costs, without drill down into specific patient diagnoses or medications. (Ketcham & Simon, 2008; Raebel, Delate, Ellis, & Bayliss, 2008) An occasional study has examined specific medications in relationship to patient needs, particularly investigating the problem of non-covered medications, mostly related to psychotropic medications. (H. Chen, et al., 2008; Zivin, McCammon, Davis, Halasyamani, & Kales, 2008) In the absence of such data, studies have relied upon the perception of results rather than the results themselves. (Epstein, Rathore, Alexander, & Ketcham, 2008; A. V. Law, Okamoto, & Brock, 2008)

We want to see if cost has shifted to prescriptions from previous reliance upon other types of care. In order to do this, we must identify Medicare patients within the dataset from the year 2005