Chapter 5
Extracting Data from the National Inpatient Sample

INTRODUCTION

In the other type of health care database that we discuss in this chapter, there are multiple columns for each patient observation. It is more difficult to find both the most frequently occurring codes, or to find patients with specific codes for the purpose of extraction. For this reason, many studies focus on the primary diagnosis or procedure. We will provide the programming necessary to find the most frequent codes and to find the patients who have a specific condition.

Another aspect of preprocessing we will explore in this chapter using the National Inpatient Sample is that of propensity scoring. When it is not possible to perform a randomized, controlled trial, an attempt is made to emulate such a trial by comparing two observational subgroups. The two groups are matched based upon demographic factors and related patient conditions. It is possible to define a level of patient severity and then to match patients with the severity level as part of the propensity score.

BACKGROUND

The National Inpatient Sample is frequently used to examine health outcomes. While it is very complete concerning inpatient events, it is limited in that there is no longitudinal follow up. For this reason, the National Inpatient Sample is used primarily to examine national trends and patterns.

DOI: 10.4018/978-1-61520-905-7.ch005
The National Inpatient Sample (NIS) also contains information identified by geographic region and by hospital as well as by hospital type. It contains several patient severity scores that can be used to create a propensity score for matched samples. In addition, it contains detailed information concerning the patient condition and procedures performed while in the hospital. There are up to fifteen columns available to list ICD9 diagnosis codes and another fifteen columns to list ICD9 procedure codes. These columns can be used to define a propensity code. They can also be used to extract subsamples with specific conditions or procedures performed.

Propensity scores are used for many reasons. The most common is to define patient risk scores for death, for disease progression, or for heart disease risk. (Ankle Brachial Index, et al., 2008; Beale, et al., 2008; Eichler, et al.; Karim, et al., 2008; Kulkarni, et al., 2007; May, et al., 2009; Pulitano, et al., 2007; Qin, et al., 2008; Tleyjeh, et al., 2008) However, they are also used to perform case-matching. (Austin & Austin, 2007, 2008b; Falagas, Mourtzoukou, Ntziora, Peppas, & Rafaillidis, 2008) Unfortunately, many studies do not provide information as to how propensity is defined. Moreover, there are often errors in the definition of such a propensity score. (Austin, 2008; Blackstone, 2002; Rosenbaum & Rubin, 1984) These errors can give results that are highly misleading. (Austin, Grootendorst, & Anderson, 2006; Austin, Grootendorst, Normand, & Anderson, 2006) The National Inpatient Sample contains several severity scores that can be used for comparison purposes.

One of the biggest problems is that any propensity score should match the severity of the patient’s condition in addition to the basic demographic factors. However, such matching depends upon how the severity level is defined, which often depends upon a separate propensity (or severity) score. As we will explain in detail in Chapter 11, this is no trivial matter. (Patricia B Cerrito, 2009b)

Because of its size, the National Inpatient Sample can be used to investigate rare occurrences. It does not have actual reimbursement information, containing information on hospital charges only. However, it does have a breakdown by hospitals, many of which are identified by name, and it has information on the hospital’s case-mix ratio. The term, case mix, refers to the type or mix of patients treated by a hospital or unit. Various funding models reimburse hospitals based upon this case mix, and a higher case mix ratio indicates more severe patients and higher reimbursements. (Brock, et al., 2007; R. H. Mehta, et al., 2008; Nicholl & Nicholl, 2007; Segal & Whyte, 1997; Stineman, 1995; VanBibber, et al., 2006) We can, however, use this case mix value to examine the relationship to the severity indices and to the patient conditions to see if there are hospitals “overcoding”, or providers who overstate the patient’s level of severity. We can also see if some providers are understating these conditions.

**PREPROCESSING DATA FROM THE NATIONAL INPATIENT SAMPLE**

The National Inpatient Sample has up to 15 columns of diagnosis codes and another 15 columns of procedure codes. The first step is to combine all 15 columns into one column so that a frequency count can be taken. Then, we can see the most common diagnosis or procedure codes in a subgroup of the NIS. We look at the first ten codes.

```plaintext
data work.removeblanks;
set work.NISdataset;
if (pr1='invl') then pr1=' ';
if (pr2='invl') then pr2=' ';
if (pr3='invl') then pr3=' ';
```