

Association for Information Systems AIS Electronic Library (AISeL)

AMCIS 2001 Proceedings

Americas Conference on Information Systems
(AMCIS)

December 2001

Using Data Mining Techniques to Characterize Diagnostic and Procedural Patterns

William Spangler
Duquesne University

Jerrold May
University of Pittsburgh

David Strum
Queen

Luis Vargas
University of Pittsburgh

Follow this and additional works at: <http://aisel.aisnet.org/amcis2001>

Recommended Citation

Spangler, William; May, Jerrold; Strum, David; and Vargas, Luis, "Using Data Mining Techniques to Characterize Diagnostic and Procedural Patterns" (2001). *AMCIS 2001 Proceedings*. 73.
<http://aisel.aisnet.org/amcis2001/73>

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in AMCIS 2001 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

USING DATA MINING TECHNIQUES TO CHARACTERIZE DIAGNOSTIC AND PROCEDURAL PATTERNS

William E. Spangler
Duquesne University
spangler@duq.edu

Jerrold H. May
University of Pittsburgh
jerryamay@pitt.edu

David P. Strum
Queen's University
strumd@post.queensu.ca

Luis G. Vargas
University of Pittsburgh
vargas@katz.pitt.edu

Abstract

This research explores the use of data mining techniques to identify diagnostic and procedure code usage patterns in two American hospitals. To determine code utilization, we borrow measures of firm concentration from industrial economics. We then use those measures (1) to ascertain the extent to which physicians utilize the available codes in classifying patients, and (2) to discover the nature of the associations between diagnostic and procedure codes. By characterizing associations between diagnostic and procedure codes in more general terms as relationships between problems and solutions, we seek eventually to extend the use of these techniques to other domains.

Introduction

In the medical domain, there are two generally accepted hierarchical numerically coded knowledge bases. One is used for diagnoses and the other for surgical and medical procedures. This research explores the use of data mining techniques to identify diagnostic and procedure code usage patterns in two American hospitals. To determine code utilization, we borrow measures of industry concentration from industrial economics. We then use those measures (1) to ascertain the extent to which physicians utilize the available codes in classifying patients, and (2) to discover the nature of the associations between diagnostic and procedure codes. By characterizing associations between diagnostic and procedure codes in more general terms as relationships between the descriptions of problems and descriptions of the solutions to those problems, we seek eventually to extend the use of these techniques to other domains.

This paper focuses on the domain-specific issues inherent in discovered code patterns, particularly the incidence of diagnostic codes, and the type and number of procedural codes associated with each. Understanding the circumstances under which diagnostic and procedural codes are used can assist various types of decision makers in the medical domain. For example, models of code usage can assist hospital managers in scheduling patients (by allowing predictions of medical procedures from initial diagnoses), and can assist insurance and other types of auditors (by conversely allowing predictions of expected diagnoses from observed procedures). While specific code patterns are primarily of interest to the medical community, the notion of diagnostic/procedural code relationships potentially is applicable in a more general sense to any domain—and data set—that can be described in terms of relationships between problems (diagnoses) and solutions (procedures).

Impacts on Code Usage Patterns

A number of factors potentially affect the relative incidence and types of medical codes used by physicians in hospitals, including the type of hospital (e.g., primary, secondary, tertiary care hospitals, teaching hospitals, and trauma centers), the nature of the particular specialty (such as surgery), rules and regulations prescribing certain types and standards of care, as well as the cultural and regulatory environment of various nations. Stausberg et al, for example, studied diagnostic and procedural code usage in a

German teaching hospital, and discovered that a small number of diagnostic and procedure codes tended to predominate across cases (i.e., a small number of the codes tended to account for a majority of the observed cases) (Stausberg, et al., 2001). This led Stausberg et al to argue in particular for the replacement of the ICD coding scheme for the classification of morbidity. Our research seeks, in part, to determine whether the same usage patterns are evident in American hospitals, and regardless of the outcome, to discover what other factors might be involved in code utilization. The Stausberg et al findings might be generalizable across nations and hospitals, but they also might be attributable to German cultural and regulatory practices, the type of hospital and specialties practiced, and so on.

Research Questions

The specific research questions under investigation in this research are as follows.

- Which factors determine or impact code usage patterns? As noted, potential factors include culture, hospital type, specialty type, and regulatory environment. They also include potential interactions between and among factors. For example, any differences between our study and the Stausberg et al study (i.e., a cross-national difference) hypothetically could be explained by the differences in hospital types, specialties practiced, individual physicians, and/or regulations.
- How do those factors impact the various aspects of code usage patterns? Code usage patterns include a) the types of codes used, b) the relative incidence or frequency of particular codes, and c) the specific relationships between diagnostic and procedure codes (i.e., the types and relative frequencies of procedure codes associated with each diagnostic code)

Research Method and Data Set

The empirical data were collected from two independent hospitals. The first ('PCH') is a primary/secondary care hospital and trauma center, while the second ('TCH') is a tertiary care and teaching hospital. The data include 59864 separate cases (patient surgeries), each containing 23 attributes detailing the diagnoses, procedures, patient demographic information, and information about the individual surgeon and anesthesiologist. The relative incidence of diagnostic and procedure codes, and the relationships between them, is characterized by the frequency distribution patterns of procedures, diagnoses, and procedures within diagnostic categories. In order to compare the various distribution patterns across potentially hundreds of different code categories, we can quantify the patterns in a single number called a *concentration index (CI)*. A concentration index is a concept borrowed from the field of industrial economics, which attempts to characterize the concentration (and dominance) of various firms within a particular industry (Shepherd, 1997). Because we are attempting to characterize code usage concentration, as well as the concentration of procedure codes within particular diagnostic categories, the use of concentration indices in the ostensibly unrelated field of data mining (and medicine) seems appropriate.

The degree of concentration can be measured using various methods. For the initial analysis of the data, we chose the Hirschman-Herfindahl index (HHI), which serves as an example of these types of methods. The HH method calculates an index from the following formula:

$$\sum_{i=1}^n p_i^2$$

Where:

n = number of firms

P_i is the percentage share of the i^{th} firm (* 100) -- ($i = 1 \dots n$).

The output of the formula is an HHI number between 0 and 10,000. Specifically, larger numbers indicate more highly concentrated industries (where one or a few firms dominate), while smaller numbers indicate more diverse industries (where the market is shared somewhat equally by a number of firms). In the domain of diagnostic and procedure codes, the industry concept is replaced by a particular medical context (e.g., hospital or specialty) and the firms are replaced by the codes themselves. In this regard, an HHI indicates the extent to which diagnoses and procedures are represented in a few, or perhaps across many, individual codes. For example, a large HHI number (generally over 1000) would indicate that a relatively small number of diagnostic codes accounts for a disproportionately large number of all diagnoses. An HHI also can be used to measure the concentration of procedure codes within individual diagnostic codes. In this case, each diagnostic code (industry) will have associated procedure codes (firms), and the codes will be variously concentrated within individual diagnoses, as calculated by the HHI.

Preliminary Results and Discussion

Preliminary indications of code usage are somewhat limited at this point, and are presented here mainly as an illustration of the types of usage patterns that might be ascertained over the course of this research. Results are described in two forms: (1) the percentage of procedure and diagnostic codes used (out of the total available), by hospital; and (2) the HHI concentration for procedure and diagnostic codes, again by hospital. In the first case, the data indicate that utilization of procedure codes in both hospitals is close to 50%, with the utilization for hospital TCH (52.7%) being somewhat higher than hospital PCH (43%). The average for both hospitals is 48.3%. By contrast, diagnostic code utilization is much lower, with TCH using only 20% of all diagnostic codes and PCH using 18.9%. The primary initial conclusion is that code utilization varies based on the type of code, with physicians using a much larger percentage of available procedure codes than diagnostic codes.

For the concentration index (HHI), the results also vary depending on the type of code. The overall concentration of procedure codes, as measured by the HHI, is considered low, with the average HHI for the two hospitals averaging 493. The HHI did not vary significantly between the two hospitals (PCH = 494; TCH = 492). The average HHI for the diagnostic codes was somewhat higher (565), but varied considerably between the two hospitals (PCH = 355; TCH = 741). The difference between diagnostic and procedure codes is attributable to the relatively higher (although still low) concentration for hospital TCH. Here, the initial conclusion is that, although not all of the codes are used (as indicated above), the concentration of the codes actually used is low (for both procedure and diagnostic codes). That is, code usage is reasonably well distributed, and is not dominated by a small group of codes.

Future Work

Finishing this research entails continuing the analysis of code usage patterns described above, which will include:

- applying the utilization and concentration analyses to the more specific hospital and specialty descriptors—for example, to determine whether there are differences between generic categories of diagnostic and procedure codes, or between various specialties
- extending the analysis to determine the relationships between diagnostic and procedure codes
- comparing the results to other studies of code usage, including those in other nations, in an attempt to understand what types of differences might exist, and why

In each case, this includes the continued use of HHI in characterizing code usage, which in turn makes this research potentially generalizable outside of the medical domain. A concise measurement of problem and solution characterizations (i.e., codes) should provide insight into the generic question of how decision makers approach the task of formulating problems and constructing plans for their solution.

References

- Shepherd, W.G. *The Economics of Industrial Organization*, Upper Saddle River, NJ: Prentice Hall, 1997.
- Stausberg, J., Lang, H., Obertacke, U. and Rauhut, F. "Classifications in Routine Use: Lessons from ICD-9 and ICPM in Surgical Practice," *Journal of the American Medical Informatics Association* (8:1), 2001, pp. 92-100.