Proceedings of the 6th INFORMS Workshop on Data Mining and Health Informatics (DM-HI 2011)
P. Qian, Y. Zhou, C. Rudin, eds.

Behavior of Liver Disease Diagnoses in a Population-Based Healthcare Insurance Claims Database

Margrét V. Bjarnadóttir
Robert H. Smith School of Business
University of Maryland
College Park, MD

Yihan Guan
Department of Management Science and Engineering
Stanford University
Stanford, CA
yihan@stanford.edu

Abstract

Drug-induced toxicities are key considerations for drug classes treating chronic conditions, but knowledge about long-term pharmacological and toxicological effects from traditional drug studies is often limited. Since drug hepatotoxicity is an important cause of liver diseases and a common reason for drug withdrawals, we characterize the behavior of liver disease diagnoses in a population-based claims database, as a building block for studying long-term drug toxicities using claims data. Our analysis show that (1) common liver diagnoses are correlated; (2) liver diagnoses evolve as sequences in one's claims history, and most sequences consist of no more than three diagnoses; (3) a long time period is needed to detect sequences of rare diagnoses; (4) those who initiate liver diseases later tend to develop diagnosis sequences faster. Furthermore, our analysis demonstrates that "one test fits all" may not be the best surveillance design, rather the statistical testing needs to be tailored to the behavior of different diagnosis codes, as they represent different conditions developed in distinct ways. We conclude that characterizing the behavior of diagnoses of liver diseases in claims data yields valuable insights to the design of an effective long-term drug safety surveillance.

Keywords: Diagnosis sequence, drug-induced hepatotoxicity, liver diseases, healthcare insurance claims data

Introduction

Even if medications have been proven safe and effective, knowledge about long-term pharmacological and toxicological effects is often limited. Traditional drug studies are short-term clinical trials, thus long-term benefits and risks have not been fully determined from these studies. On the other hand, long-term drug-induced toxicities are key considerations for drug classes treating chronic conditions, because exposure to these drugs can be years or decades. In addition, recent studies have suggested that it may be cost-effective to administer low-cost statins to most persons with modestly elevated cholesterol or any coronary heart disease risk factors as a primary prevention strategy [4]. With a transition from strictly therapeutic drug use to a preventative use, obtaining knowledge of long-term effects of drugs is more urgent than ever.

The liver plays a central role in transforming and clearing chemicals from the bloodstream, and is therefore highly susceptible to drug-induced toxicity. Drug-induced hepatotoxicity is a leading cause of liver diseases and a common reason for drug withdrawals or black box warnings. On the other hand, drug-induced liver injury remains a diagnostic challenge for physicians, because it might have similar characteristics of acute and chronic liver diseases [5]. As a result, liver disease is a natural starting point to study long-term drug effects.

One possible data source to study long-term drug effects is healthcare insurance claims, which has the advantage of a large volume, enabling the study of rare outcomes. In this paper, we seek to understand and characterize the behavior of diagnoses of liver diseases in a population-based claims database, as a building block for studying long-term

drug toxicities using claims data.

Claims data and diagnosis coding structure
Claims data are generated when healthcare providers send claims to insurance companies or third parties to receive reimbursements for their services. Claims data reflect both medical and pharmacy claims. Medical claims include for example, diagnosis codes, treatment codes for any procedures performed, date and place of service, and associated cost. In claims data, diagnosis are coded using International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). An ICD-9-CM code consists of a three-digit category code plus additional fourth and fifth digits to describe conditions in further details. The description of a four-digit code is therefore more specific than that of its three-digit "super" code. As an example, 070 represents viral hepatitis while 070.0 stands for viral hepatitis A with hepatic coma. In this study, we analyze liver diseases represented by groups of ICD-9-CM codes. Table 1 lists the coding groups, where the three-digit codes include all the subsequent four- or five-digit codes.

Diagnosis coding group	Description
070	Viral hepatitis
155	Malignant neoplasm of liver, primary
197.7	Malignant neoplasm of liver, secondary
570	Acute and subacute necrosis of liver
571	Chronic liver disease and cirrhosis
572	Liver abscess and sequelae of chronic liver diseases
573	Other disorders of liver
794.8	Abnormal results of function study of liver
864	Injury to liver

Table 1: Liver disease diagnosis coding groups analyzed in the study.

Behavior of liver disease diagnoses

Study population
This study utilizes a longitudinal claims database developed by VeRisk Health Inc.[1] The study population includes members that have at least four consecutive years of claims history starting from the date of their first claim in the database and at least one pharmacy claim. This inclusion criteria results in a population of 487,789 members. Among these members, 41,283 have at least one claim that includes a diagnosis of liver diseases.

Correlations
Traditional multiple hypotheses testing methods assume independence of the hypotheses being tested[1] and hence the resulting adjustment methods for multiple testing such as Bonferroni correction and Šidák correction are not appropriate for dependent hypotheses. To mitigate the risk of producing excessive false positives or overlooking true positive signals, permutation-based methods to adjusting for dependence in genome-wide association studies have been proposed [2, 3]. These methods can be extended to drug safety studies.

A successful drug safety surveillance system needs to monitor multiple adverse events, and hence one needs to understand correlations between adverse events to ensure statistical validity and precision. In the case of liver diseases, it is likely that several diagnoses are found in the claims of a single patient within a month or even on a single day. Below we analyze the correlations between different diagnosis codes from all members who have liver diagnoses. Table 2 lists the diagnosis pairs with the highest correlation coefficients. There is a high correlation (0.353) between primary liver cancer (155) and secondary liver cancer (197.7), a moderate correlation (0.104) between acute and subacute necrosis of liver (570) and liver abscess and sequelae of chronic liver diseases (572) as well as between liver abscess and sequelae of chronic liver diseases (572) and primary liver cancer (155). These correlations are not surprising as when a cancer is first discovered in the liver, it is often not clear if it is primary or secondary. Through biopsy or other procedures the nature of the cancer may be identified, and the diagnosis is updated accordingly. Acute and subacute necrosis of the liver (570) are often results of toxicity (e.g., mushroom, alcohol, or drugs). Sequelae of

[1] Verisk Health Corporate, 130 Turner Street Seventh Floor, Waltham, MA 02453. http://www.veriskhealth.com/.

Diagnosis	Diagnosis	Correlation
Malignant neoplasm of liver, primary (155)	Malignant neoplasm of liver, secondary (197.7)	0.353
Acute&subacute necrosis of liver (570)	Liver abscess&sequelae of chronic liver diseases (572)	0.104
Liver abscess&sequelae of chronic liver diseases (572)	Malignant neoplasm of liver, primary (155)	0.104
Other disorders of liver (573)	Malignant neoplasm of liver, primary (155)	0.067
Other disorders of liver (573)	Malignant neoplasm of liver, secondary (197.7)	0.051

Table 2: Most correlated liver diagnoses and correlation coefficients.

chronic liver disease (572) can be a consequence of acute or subacute necrosis from which a patient does not fully recover. Finally, it is known that any chronic liver disease is associated with an increased risk of liver cancer, hence a correlation between sequelae of chronic liver disease and liver cancer is unsurprising.

Patterns in diagnosis coding sequences

A natural question to ask is how do a member's liver diagnoses evolve in his/her claims history? To answer this question, we first look at the number of diagnosis codes that members have during at least four consecutive years. Notice that consecutive repeats of a code is regarded as one code, whereas returning to one code from a distinct code is regarded as a different code in a sequence. Again we focus on members having at least one liver code in their claim histories. We find that 78.1% of the population have only one code, 93.5% have at most three, and 1.9% have more than seven. As an example of coding sequences, Figure 1 plots a code sequence of a member who suffered chronic liver diseases. This member experienced nine changes in liver diagnoses during 3.3 years, switching between abnormal results of a liver scan (794.8), other disorders of liver (573), and chronic liver disease and cirrhosis (571).

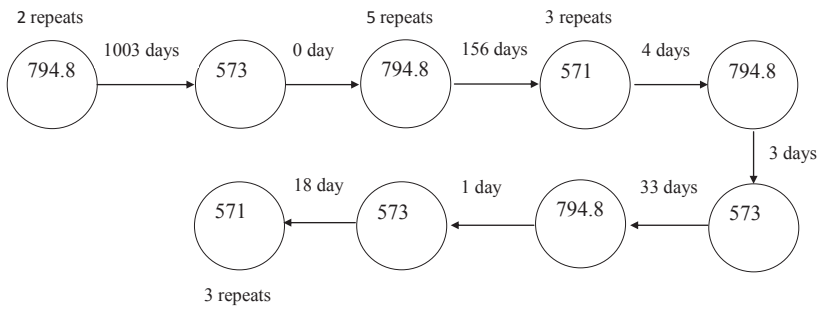

Figure 1: Sample sequence of liver diagnosis codes. 794.8: abnormal liver scan, 573: other disorders of liver, 571: chronic liver disease and cirrhosis.

Next we consider how the diagnosis codes are distributed in the study population. Here each code is counted only once if a member ever has this code despite of the fact that each code can repeat or be recurrent. For example, the member shown in Figure 1 has three codes: 794.8, 573, and 571. We find that codes 794.8, 571, and 573 are the most frequent diagnoses, found in 38.9%, 36.9%, and 36.2% of the population, respectively. Next are codes 070 and 572, experienced by 10.4% and 2.5% of the population, respectively. Each of codes 155, 197.7, 570, and 864 is found in less than 2% of the population, and 4.7% in the aggregate.

Now we analyze the most common sequences of codes. Since 93.5% of the population have diagnosis sequences of three codes at most (returning to one code from a distinct code is regarded as a different code), we focus on analyzing sequences of no more than three codes. Denote the ith diagnoses in a sequence by x_i, $i = 1, 2, 3.....$ If a sequence has only one code, $x_2 = \emptyset$. Similarly, if a sequence has only two codes, $x_3 = \emptyset$. Table 3 lists the most common single

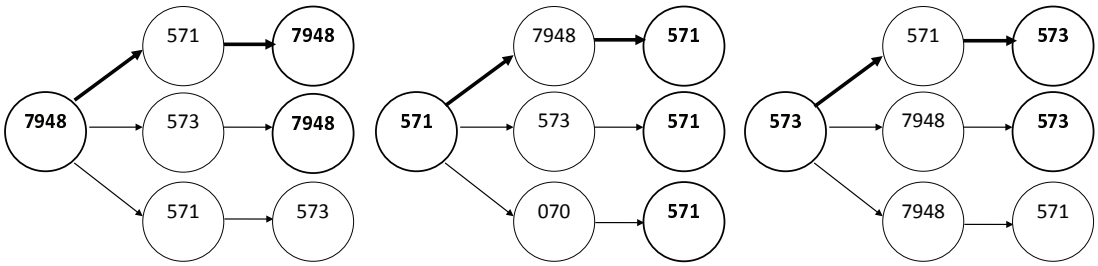

Figure 2: Common 3-code sequences. The most frequent sequence starting from a first code is shown in the top row.

	$P(x_1)$		$P(x_1\|x_2)$		$P(x_3\|x_1, x_2)$
x_1 = 794.8	0.33	$x_2 = \emptyset$	0.8		
		x_2= 571	0.12	$x_3 = \emptyset$	0.54
				x_3=7948	0.33
x_1 = 571	0.28	$x_2 = \emptyset$	0.8		
		x_2= 794.8	0.08	$x_3 = \emptyset$	0.56
				x_3=571	0.37
x_1 = 573	0.28	$x_2 = \emptyset$	0.8		
		x_2= 794.8	0.08	$x_3 = \emptyset$	0.59
				x_3=573	0.24

Table 3: Common single diagnoses and sequences of two and three codes.

diagnoses and sequences of two and three codes, and the corresponding conditional probabilities. For instance, the most common single diagnosis is abnormal liver scan (794.8), followed by chronic liver disease and cirrhosis (571) and other disorders of liver (573); the most common 2-code sequence is abnormal liver scan→chronic liver disease and cirrhosis (x_1=7948, x_2=571); the most common 3-code sequence is abnormal liver scan→chronic liver disease and cirrhosis→abnormal liver scan (x_1=7948, x_2=571, x_3=7948). It can be interpreted as that 33% of the members who ever have liver codes start with an abnormal liver scan, and 80% of these members are not diagnosed subsequently whereas 12% of them are diagnosed as chronic liver disease and cirrhosis. Not surprisingly, having multiple abnormal liver scan is quite common once a member is diagnosed as chronic liver disease and cirrhosis or other disorders of liver. Figure 2 shows common three-code sequences. Interestingly, a three-code sequence is more likely to traverse back to the starting code than to arrive at a new code.

In addition to sequences, time between codes also plays an important role in understanding the behavior of liver diagnoses. Particularly, as our ultimate objective is to design an effective long-term drug safety surveillance, how long a typical progression of liver disease takes (reflected as the duration of a sequence) is an essential factor to consider when we design an observation horizon. Therefore, we study the distribution of days between different diagnoses. Table 4 lists the mean, the median and the top 5 percentile value of time between x_1 and x_2, x_1 and x_3, x_1 and x_5, x_1 and x_{10}, and x_1 and x_{20}. For example, a 3-code sequence on average can be detected in 425 days (1.2 years) after the onset of the first code. 95% of 3-code sequences can be observed in 1500 days (4.1 years) after the onset of the first code, whereas 95% of 20-code sequences need 2154 days (5.9 years) to be observed. Note that these estimates are affected by members' individual membership duration in the database.

Rare diagnoses

Other than common diagnoses, rare diagnoses (e.g., liver cancer) are of great importance when the target is to detect severe or fatal adverse drug events. Define the duration of a liver diagnosis sequence as the time between a member's first diagnosis (not limited to a liver diagnosis) in the database and the last liver diagnosis. Table 5 compares the average duration of the sequences containing primary liver cancer (155) or secondary liver cancer (197.7) to that of the sequences not containing either cancer code. For short sequences (length=2, 3, or 5), the ones containing cancer

	Mean	Median	Top 5%
x_1 to x_2	321	78	1365
x_1 to x_3	425	195	1500
x_1 to x_5	621	441	1715
x_1 to x_{10}	885	845	1869
x_1 to x_{20}	1048	1026	2154

Table 4: Descriptive statistics of time (days) between diagnoses.

Sequence length	Containing 155	Containing 197.7	No 155 or 197.7
2	1511	1940	1326
3	1595	1598	1346
5	1718	1772	1516
10	1641	1754	1755
>10	1688	1728	1794

Table 5: Average duration (days) of sequences containing primary liver cancer (155), secondary liver cancer (197.7), or neither.

codes have longer durations on average than those not containing cancer codes; however for longer sequences (length $\geq$10), the average durations of these two types of sequences become close, both reaching 5 years. Moreover, time until the first incidence of common liver codes (e.g., 571, and 794.8) and the rare liver codes (e.g., 155 and 197.7) differs significantly on average. Table 6 lists time to the first incidence of codes 571, 794.8, 155, and 197.7. It takes one more year (370 days) on average to observe the first incidence of code 155 than to observe that of code 794.8. Here we evaluate the time until the first incidence of codes in two cases. The first case is that all members having liver diagnoses are included (n=41,283). The second case is that those who have liver codes during the first year of their claims history are excluded (n=31,729). These excluded members may represent a group with pre-established liver conditions. In the second case (numbers are shown in parentheses in Table 6), time to the first incidence of a liver code is therefore longer than that in the first case.

The above observations suggest that when choosing an observation horizon for a long-term drug safety surveillance, one needs to carefully consider what types of signals are followed. If traditional "testing each code separately" approach is taken, or if there is a pre-period in the testing design that eliminates members with any liver pre-conditions, 3 years may not be an enough time window to detect rare diagnoses which are often (but not always) results of other conditions. If the target is to detect an increased risk of more common sequences, 2 to 3 years of observation could be sufficient. On the other hand, if one is interested in sequences of more than three diagnoses, 4 years or longer period is necessary.

Comparing diagnosis sequences in two periods

In order to determine whether drug-induced toxicities accumulate over time, we need to compare outcomes in disjoint time intervals. In this section, we compare the behavior of liver code sequences in two observation periods.

Definition of periods
We define three consecutive periods for each member starting from the date of the member's first claim in the database:

Diagnosis code	Description	Mean	Median	Top 5%
155	Primary liver cancer	1391(1572)	1505 (1589)	2397 (2431)
197.7	Secondary liver cancer	2458 (2659)	1549 (1593)	2395 (2411)
571	Chronic liver disease and cirrhosis	1153 (1390)	1077 (1293)	2242 (2322)
794.8	Abnormal liver scan	1021 (1295)	898 (1162)	2148 (2258)

Table 6: Time (days) to first liver diagnosis. Numbers in parenthesis are from members who do not have liver diagnoses during the first year of their claims histories.

a one-year pre-study period, a two-year Period I, and a two-year Period II, to contrast the behavior of diagnosis codes during Period I and II. The pre-study period is used to exclude any members with pre-established liver conditions. This criteria eliminates 5,580 members from 27,434 members who have at least one liver diagnosis and at least five consecutive years of claims history, and hence the study cohort consists of 21,854 members. In addition, only first incidence of a sequence is considered in each period.

Results
We find that the most common diagnoses during Period I and II are identical, and the population fractions of individual diagnoses are similar, except that there are 5% more incidences of code 794.8 in Period I than in Period II, while Period II has 2% more incidence of codes 573 and 571. Furthermore, the frequent 3-code sequences in the two periods agree with the ones listed in Table 3. However, average time between diagnoses are shorter in Period II than that in Period I. For instance, average time between the first and the third codes in period II is 96 days (median: 42 days), compared to 118 days (median: 59 days) in Period I. Sequences in Period II run faster than those of same lengths in Period I. For example, sequences of five codes in period II run for 169 days on average, from the onset of the first liver code (median: 131 days), while those in Period I run for 249 days on average (median: 231 days). Considering members in Period II have no liver condition in Period I, these results indicate that with a longer exposure to a potential source of hepatotoxicity, those who initiate liver diseases later in fact develop a diagnosis sequence faster. It is worth noting that because the database used in this study consists of claims from a population having at least one pharmacy claim, the underlying characteristics of the study cohort might not be ideally representative of a general population. In addition, we have not accounted for aging effects. A further population-wide study is needed to validate these observations. The current results encourage more in-depth analyses of diagnosis sequences and their time durations for monitoring drug effects on liver disease progression.

Discussion

Obtaining knowledge of long-term effects of drugs is an urgent task for today's drug safety surveillance. In this study, we characterize the behavior of diagnoses of liver diseases in claims data, as a fundamental step to investigate long-term drug toxicities . Our results show that common liver diagnoses are correlated; liver diagnoses evolve as sequences in one's claims history, and most sequences consist of no more than three diagnoses; a longer time period is needed to detect sequences of rare diagnoses; those who initiate liver diseases later tend to develop diagnosis sequences faster, when we compare two periods of equal lengths. Furthermore, our analysis clearly shows that "one test fits all" may not be the best surveillance design, rather the statistical testing needs to be tailored to the behavior of different diagnosis codes, as they represent different conditions developed in distinct ways. Our next steps include validating the current observations in a larger population-wide data set, developing and testing appropriate signals to detect increased risk in long-term drug studies while maintaining statistical power. We conclude that understanding the behavior of diagnoses of liver diseases in claims data offers valuable insights to the design of an effective long-term drug safety surveillance.

References

[1] A. Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4):347–388, Aug 2008.

[2] X. Gao, J. Starmer, and E. R. Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32(4):361–369, 2008.

[3] B. Han and E. Eskin. Multiple testing in genetic epidemiology. In *Encyclopedia of Life Science*. John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0022497, 2010.

[4] L. D. Lazar, M. J. Pletcher, P. G. Coxson, K. Bibbins-Domingo, and L. Goldman. Cost-effectiveness of statin therapy for primary prevention in a low-cost statin era. *Circulation*, 124(2):146–153, July 2011.

[5] R. A. Nathwani and N. Kaplowitz. Drug hepatotoxicity. *Clinics in Liver Disease*, 10(2):207–217, May 2006.