Application of Information-Theoretic Data Mining Techniques in a National Ambulatory Practice Outcomes Research Network

Adam Wright[a], Thomas N. Ricciardi[a,b], and Martin Zwick[c]

[a] *Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland Oregon USA*

[b] *GE Healthcare Technologies, Waukesha Wisconsin, USA*

[c] *Systems Science Ph.D. Program, Portland State University, Portland Oregon USA*

Abstract

The Medical Quality Improvement Consortium data warehouse contains de-identified data on more than 3.6 million patients including their problem lists, test results, procedures and medication lists. This study uses reconstructability analysis, an information-theoretic data mining technique, on the MQIC data warehouse to empirically identify risk factors for various complications of diabetes including myocardial infarction and microalbuminuria. The risk factors identified match those risk factors identified in the literature, demonstrating the utility of the MQIC data warehouse for outcomes research, and RA as a technique for mining clinical data warehouses.

Keywords

Data Warehouse, Data Mining, Reconstructability Analysis, Ambulatory Electronic Medical Records

Introduction

Electronic Medical Records (EMRs) are becoming increasingly important sources of clinical information for outcomes research and quality measurement. Historically, such research has been challenging because of the difficulty of aggregating EMR data across multiple institutions and providers. The Medical Quality Improvement Consortium (MQIC) helps solve that problem. MQIC is a network of 74 healthcare delivery organizations across the United States that use an ambulatory electronic health record (Centricity Physician Office, GE Healthcare Information Technologies, Waukesha, WI). MQIC member offices transmit de-identified streams of clinical data to a centralized data warehouse managed by GE. The transmission process is designed to fully respect patient privacy and comply with all regulatory requirements[1].

MQIC uses the data warehouse to accomplish its dual goals:

1. To provide summary reports about quality and patterns of care to each member office.
2. To make data available to consortium members and carefully vetted researchers for use in outcomes studies.

This study applies reconstructability analysis, or RA, to the MQIC data set. RA is a modeling technique, developed in the systems community, based in information theory and graph theory [2-8], which resembles log-linear modeling [9, 10] and Bayesian network modeling. RA is well-suited for explanatory modeling, and, compared to other techniques, is especially capable in detecting unknown nonlinear relationships and interaction effects. It can model both problems where "independent variables" (inputs) and "dependent variables" (outputs) are distinguished (called *directed systems*) and where this distinction is not made (*neutral systems*).

It is easiest to introduce RA with an example. Assume we have a directed system with several inputs A, B, C and D representing risk factors and one output Z representing a disease state. Our goal is to figure out whether A, B, C and D can be used to predict Z.

The most complete model we could form would consider all four variables jointly as predictors of Z. This model makes maximum use of the frequency distribution involving these variables so it retains all of the data's explanatory power, but it is also maximally complex and may overfit the data. At the opposite end of the spectrum is the null model, which does not use any of A, B, C or D to predict Z. This model is minimally complex, but does not attempt to explain Z.

RA attempts to find the best explanatory model along this continuum. There is a tradeoff between maximizing goodness of fit and keeping complexity

to a manageable level. Various criteria are available to help strike this balance.

In our example, RA might recommend a model that uses only B and D to predict Z – omitting A and C from the predicting relationship.

Since RA is an information theoretic technique, it measures explanatory power in terms of information, or equivalently, in terms of reduction of uncertainty about the output variable. The explanatory power of any model is defined to be the fractional reduction of uncertainty in the output achieved by the predicting components of the model. Uncertainty is the analog of variance for nominal variables. The null model discussed above, which doesn't use any of the inputs to predict the output contains 0% information and has 0% uncertainty reduction. The best model for a dataset is the one which reduces uncertainty the most while still being statistically significant. Statistical significance is evaluated using the Chi-square distribution; this requires an adequate sample size, easily satisfied by the MQIC data.

Methods

For this study, a cohort was extracted from the MQIC data warehouse consisting of 50,428 adult patients with a diagnosis of either type I or type II diabetes mellitus on their problem list who had been seen by a healthcare provider in the past two years and for whom lab tests were available. A set of variables was determined based on measures commonly used for health maintenance and quality assessment programs [11]. These measures and the number of patients demonstrating each are listed in Table 1. They are the inputs and outputs that will be used to form models.

Unlike many data mining techniques reconstructability analysis is performed on discrete, rather than continuous data. Information about problems and drugs is already discrete and binary (diagnoses or medications are either present and active in the patient's chart or not). But many of the observed variables, particularly lab tests, were measured on a continuous scale. For the purposes of RA, these values were treated as discrete by placing them into bins according to widely accepted standard reference ranges, which are described in Table 1. For example, LDL cholesterol values were classified as either normal or high based on standards from the National Heart, Lung and Blood Institute [12].

Calculations were made using the RA software programs developed at Portland State University, now integrated into the package Occam (for the principle of parsimony and as an acronym for "Organizational Complexity Computation And Modeling") [13-15].

Table 1 – Characteristics of the study population

Measure	Value	Count
Gender	Female	24,548
	Male	25,880
Deceased	Yes	963
Ever married	Yes	29,726
Age	<50	10,046
	[50,60)	11,656
	[60,70)	12,615
	≥80	5,239
Hypertension diagnosis	Yes	33,427
Hyperlipidemia diagnosis	Yes	20,741
Hypercholesterolemia diagnosis	Yes	11,818
Albuminuria diagnosis	Yes	162
Diabetes medication	Yes	45,321
Beta blocker medication	Yes	15,638
Aspirin medication	Yes	23,936
ACE Inhibitor medication	Yes	23,208
Smoking med	Yes	326
Eye exam in last year	Yes	14,426
Foot exam in last year	Yes	21,770
Smoking status at last exam	current	11,926
	former	8,239
	never	17,228
	N/A	13,035
Last HgbA1c	> 9	7,049
Last blood pressure	> 140/90	17,566
Last microalbumin	> 15	21,745
Last total cholesterol	<200	36,925
	[200,240)	9,637
	≥240	4,136
Last LDL	<130	41,931
	[130,160)	5,743
	≥160	2,754
Best LDL	<130	35,318
	[130,160)	10,698
	≥160	4,412
Last HDL	<40	14,893
	≥40	35,535
Last triglyceride	<200	36,893
	[200,240)	12,580
	≥240	955
History of myocardial infarction (mi)	Yes	1,461
Microalbuminuria (malb) diagnosis	Yes	1,230
Ischemic heart disease (ihd) diagnosis	Yes	6,970
Peripheral vascular disease (pvd) diagnosis	Yes	2,027
History of cerebrovascular accident (cva)	Yes	2,156
Congestive heart failure (chf) diagnosis	Yes	2,870

Occam was used to find models that predicted common complications of diabetes (i.e. myocardial infarction and microalbuminuria). For ease of interpretation, each complication was analyzed separately and in this analysis other complications were allowed to appear as risk factors.

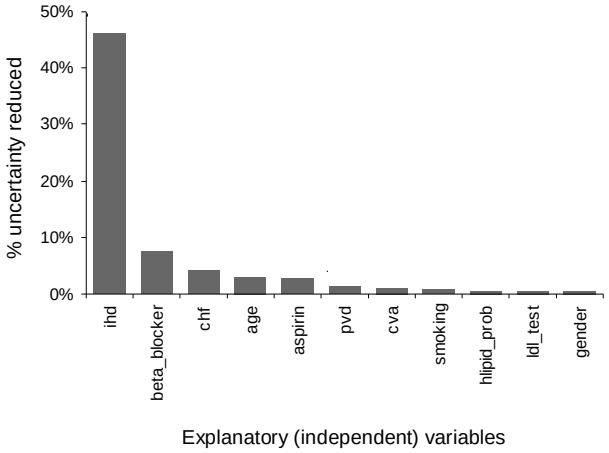
It is important to note that Occam had no *a priori* knowledge of the meaning of any variable, or the relationship between any variables. All associations found during the analysis process come directly from the data.

Results

Occam was used to generate simple (loopless) models to predict complications. For example the study examined whether or not a diabetic patient, according to their problem list, suffers from microalbuminuria, a potentially serious nephrologic complication of diabetes. Occam determined that the best predictor of microalbuminuria was the existence of an elevated urine microalbumin level in a lab test. The second and third best predictors of microalbuminuria were presence of records of a foot exam and dilated eye exam. None of the 27 other predictors in the dataset reduced uncertainty by more than 0.5%.

Occam also produced strong models for myocardial infarction (MI). The results are summarized in Figure 1. The presence of ischemic heart disease on the problem list was the single best predictor of MI. In RA terms, IHD reduced the uncertainty of having an MI on the problem list by 46.2%. Beta blockers were the next major predictor of MI. In descending order of explanatory strength congestive heart failure, old age, aspirin therapy, peripheral vascular disease, history of cerebrovascular accident, smoking, documented hyperlipidemia, best LDL on record and gender are also associated with MI in our dataset.

Figure 1 – Explanatory value of measured variables in predicting myocardial infarction.



The 19 remaining variables we considered as possible predictors each explained less than 0.5% of the total variability in MI.

The strength of each of the explanatory variables in predicting each of the six studied complications is presented in Table 2. All of these predictors were significant at the 95% confidence level, even after using the Bonferroni correction for multiple comparisons. It's important to note that explanatory strength is not necessarily symmetric –for example, IHD reduced uncertainty about MI by 46%, but MI reduces uncertainty about IHD by only 16%, as nearly everyone who has an MI also has IHD, but many people with IHD have not had an MI.

Table 2 – Explanatory strength (% uncertainty reduction) for each predictor and each studied complication

Predictor	mi	malb	ihd	pvd	cva	chf
myocardial infarct	–	0	15	1	1	3
microalbuminuria	0	–	0	0	0	0
isch. heart. disease	46	0	–	5	2	6
periph vasc disease	1	0	2	–	1	2
cerebrovasc. accid	1	0	1	1	–	1
CHF	4	0	3	3	1	–
gender	1	0	1	0	0	0
marital status	0	0	0	0	0	0
age	3	0	6	6	6	7
albuminuria	0	0	0	0	0	0
hypertension	0	0	0	1	2	0
hyperlipidemia	1	0	1	0	0	0
high cholesterol	0	1	0	0	0	0
diabetes med	0	0	0	0	0	0
beta blocker	7	0	10	2	1	6
aspirin therapy	3	0	5	1	1	0
ace inhibitor	0	0	0	0	0	0
smoking med	0	0	0	0	0	0
eye exam	0	2	0	0	0	0
foot exam	0	3	0	0	0	0
smoking status	1	0	1	2	0	1
last hgba1c	0	0	0	0	0	0
last blood pressure	0	0	0	0	0	0
last microalbumin	0	4	0	1	0	1
last cholesterol	0	0	1	0	0	0
last LDL	0	0	1	0	0	0
optimal LDL ever	1	0	2	1	0	0
near optimal LDL	0	0	1	0	0	0
last HDL	0	0	1	0	0	1
last tg	0	0	0	0	0	0

Discussion

These results are very encouraging. The best predictor of microalbuminuria was an elevated urine microalbumin lab test. Since microalbuminuria is, by definition, the presence of microalbumin in urine, this serves as a valuable control. That Occam, which had no a priori knowledge of this fundamental association, chose it as the most useful predictor, as opposed to choosing an indicator, such as marital status, for which there is no known clinical evidence of correlation is encouraging evidence of the validity of both the MQIC database and the technique of reconstructability analysis. The other two predictors that Occam found (foot exam and dilated eye exam) are most frequently performed and documented in patients with the most severe cases of diabetes, they serve as proxies for disease severity. Since the highest severity patients are most likely to suffer from microalbuminuria, we would expect the association between the two variables. These kinds of indirect associations between characteristics of a patient population would be difficult for a human to construct. Yet they often arise through reconstructability analysis, and can be extremely useful for creating models to predict diseases or detect novel interactions between diseases and treatments, particularly when direct measures of the disease states are not available in the data source. These two associations likewise serve as valuable controls in assessing the validity of the data and technique.

The results for the myocardial infarction are similarly encouraging. IHD is the major cause of MI, so its dominance in the model is expected. Further, current guidelines recommend beta blockers post-MI, so it makes sense that beta blocker use and MI would be associated. The other predictors are also well supported by the evidence. Like beta blockers, aspirin therapy is recommended post-MI by current treatment guidelines. Age and gender are known demographic risk factors for MI. Smoking and hyperlipidemia are known contributory risk factors for MI, and congestive heart failure, peripheral vascular disease and cerebrovascular accident are known to be comorbid with MI. Again, all of the selected factors are well known, and factors not known to be associated with MI risk, such as hemoglobin A1c and marital status, are excluded.

Similar results were seen for the other complications studied. For example, Occam determined that age and smoking status were risk factors for nearly all the complications, which is to be expected. It also highlighted beta blockers for the cardiac diseases (MI, IHD and CHF) and, to a lesser extent, for the vascular diseases (PVD and CVA), but not at all for the nephrologic disease (microalbuminuria) – exactly as would be expected given the indications for beta blocker therapy.

The results found in this study represent a form of machine learning: the computer learned facts about diseases in the form of associations. That the facts learned here are both clinically plausible and well supported by the literature is encouraging for two reasons. First, because it helps to validate the accuracy and usefulness of the data in the MQIC database, and second, because it demonstrates the utility of reconstructability analysis as a tool for knowledge discovery in clinical data warehouses. Automated analysis of medical records without human supervision or assistance is a difficult [16] and pressing [17] problem and the technique of reconstructability analysis appears to be a robust approach for just such analyses.

There is ample room for future work using the MQIC data warehouse and RA, both together and separately. For example, other cohorts and variables of interest could be extracted and analyzed, and more complex models with more variables, multiple predicting components or multi-variable interaction effects could be considered. It would also be useful to attempt to discover new information, instead of just confirming what is already known. For example, a researcher could build a model of the top 100 drugs, along with common causes of mortality (cancer, MI, CVA) and look for unusual links. Such a technique might be able to mine for unknown connections between drugs and conditions, such as the recently discovered link between rofecoxib (Vioxx®) and MI. Finally, at present, using Occam requires a reasonably thorough understanding of RA. Tutorial papers [5,6] and a user's manual [14] are available, but a new analysis package could also be designed which would allow non-experts to use RA by guiding the user through the process.

Summary

The MQIC data warehouse, which contains over 3.6 million detailed clinical records formatted for research and practice improvement, is described. RA, an information-theoretic technique for data mining is described and applied to the MQIC data warehouse, resulting in models which are used to predict various complications of diabetes. These risk factors match those in the literature, and have a high degree of clinical plausibility, demonstrating that:

1. Those subsets of the MQIC data warehouse used here were sufficiently accurate to ensure that the associations discovered were reasonable, and that no spurious associations were discovered.

2. RA is a sufficiently robust technique to discover valid associations in a large clinical data warehouse.

The MQIC data warehouse and RA, both singly and together are promising new tools for outcomes research and great opportunities exist for further study using them.

Acknowledgments

The authors gratefully acknowledge Tina Ho, Stuart Lopez, Sunil Luhadia, Matt Reeves, Dean Sittig and Thomas Yackel.

References

[1] Ricciardi T, Lieberman MI, Kahn MG, and Masarie FE. Clinical Terminology Support for a National Ambulatory Practice Outcomes Research Network. Proc AMIA Symp; 2005. (submitted).

[2] Ashby WR. Constraint Analysis of Many-Dimensional Relations. General Systems Yearbook. 1964;9:99-105.

[3] Klir G. Reconstructability Analysis: An Offspring of Ashby's Constraint Theory. Systems Research. 1986;3(4):267-71.

[4] Klir G, ed. Int. J. General Systems Special Issue on GSPS. 1996;24(1&2). (includes an RA bibliography).

[5] Zwick M. Wholes and Parts in General Systems Methodology. In: Wagner G, ed. The Character Concept in Evolutionary Biology. New York: Academic Press; 2001. p. 237-56.

[6] Zwick M. An Overview of Reconstructability Analysis. Kybernetes. 2004;33:877-905.

[7] Klir G. The Architecture of Systems Problem Solving. New York: Plenum Press; 1985.

[8] Krippendorff K. Information Theory: Structural Models for Qualitative Data: Quantitative Applications in the Social Sciences #62. Beverly Hills, CA: Sage; 1986.

[9] Bishop Y, Feinberg S, Holland, P. Discrete Multivariate Analysis. Cambridge, MA: MIT Press; 1978.

[10] Knoke D, Burke PJ. Log-Linear Models: Quantitative Applications in the Social Sciences Monograph # 20. Beverly Hills, CA: Sage; 1980.

[11] Joyner L, McNeeley S, Kahn R. ADA's (American Diabetes Association) provider recognition program. HMO Pract. 1997;11(4):168-70.

[12] National Heart, Lung, and Blood Institute. National Cholesterol Education Program (NCEP). Third report of the NCEP on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). NIH Publication No. 01-3305.2001.

[13] Hosseini J, Harmon RR, Zwick M. Segment Congruence Analysis Via Information Theory. In: Proceedings of International Society for General Systems Research. 1986:G62-77.

[14] Zwick M. "Discrete Multivariate Modeling". 2005. http://www.sysc.pdx.edu/res_struct.html.

[15] Willett K, Zwick M. A Software Architecture for Reconstructability Analysis. Kybernetes. 2004;33:997-1008.

[16] Sittig DF. Grand challenges in medical informatics? J Am Med Inform Assoc. 1994;1(5):412-3.

[17] Office of the National Coordinator for Health Information Technology. "Health IT Strategic Framework". 2004. http://www.hhs.gov/healthit/frameworkchapters.html.