ELSEVIER

Artificial Intelligence in Medicine 26 (2002) 37–54

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Data mining a diabetic data warehouse

Joseph L. Breault[a,b,*], Colin R. Goodall[c,d], Peter J. Fos[e,b]

[a]*Family Medicine, Ochsner Clinic Foundation, New Orleans, LA 70121, USA*
[b]*Health Systems Management, Tulane University, New Orleans, LA 70112, USA*
[c]*AT&T Shannon Research and Technology Laboratory, Middletown, NJ 07748, USA*
[d]*Adjunct Appointment, Biostatistics, Tulane University, New Orleans, LA 70112, USA*
[e]*School of Dentistry, University of Nevada Las Vegas, Las Vegas, NV 89154, USA*

Received 5 March 2002; accepted 12 March 2002

Abstract

Diabetes is a major health problem in the United States. There is a long history of diabetic registries and databases with systematically collected patient information. We examine one such diabetic data warehouse, showing a method of applying data mining techniques, and some of the data issues, analysis problems, and results. The diabetic data warehouse is from a large integrated health care system in the New Orleans area with 30,383 diabetic patients.

Methods for translating a complex relational database with time series and sequencing information to a flat file suitable for data mining are challenging. We discuss two variables in detail, a comorbidity index and the HgbA1c, a measure of glycemic control related to outcomes. We used the classification tree approach in Classification and Regression Trees (CART®) with a binary target variable of HgbA1c >9.5 and 10 predictors: age, sex, emergency department visits, office visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, retinopathy, end-stage renal disease.

Unexpectedly, the most important variable associated with bad glycemic control is younger age, not the comorbiditity index or whether patients have related diseases. If we want to target diabetics with bad HgbA1c values, the odds of finding them is 3.2 times as high in those <6.5 years of age than those older. Data mining can discover novel associations that are useful to clinicians and administrators.

Keywords: Data mining; Diabetes; Data mining software CART

1. Introduction

From the moment of birth to the signing of the death certificate, data are collected at almost every contact of each individual with providers of healthcare in the United

* Corresponding author. Tel.: +1-504-243-6021; fax: +1-504-243-6052.
E-mail addresses: joebreault@tulanealumni.net (J.L. Breault), cgoodall@att.com (C.R. Goodall), peter.fos@ccmail.nevada.edu (P.J. Fos).

0933-3657/02/$ – see front matter
PII: S0933-3657(02)00051-9

States (and many other countries). These data include administrative, demographic, health status, clinical, pharmaceutical use, and financial details. Increasingly, data are abstracted from written records or entered directly at a workstation into an extensive health information system [19].

Clinical and financial healthcare data are massive. Uniform billing data are collected nationwide for discharges—tens of millions of records per year. Clinical data collected at any given institution are massive. When the National Academy of Sciences convened a conference on Massive Data Sets in 1995, the presentation on healthcare noted that ''massive applies in several dimensions. . .the data themselves are massive, both in terms of the number of observations and also in terms of the variables. . .there are tens of thousands of indicator variables coded for each patient'' [21]. Moreover, we multiply this by the number of patients in the United States, which is virtually the same as the population, namely hundreds of millions.

Extensive diabetic registries have existed for decades, some of which have evolved into data warehouses. Data mining techniques have recently been applied to them in an attempt to predict diabetes development or high-risk cases, to find new ways to improve outcomes, and to detect provider outliers in quality of care or in billing services [9,23,25,32,33]. We examine one such diabetic data warehouse, showing a method of applying data mining techniques, and some of the data issues and analysis problems.

2. Understanding the medical problem domain

The population of diabetic patients is important in healthcare for a number of reasons. It is large with 15.7 million people in 1998 or 5.9% of the population of the United States having diabetes, 8.2% of those 20 and older, and 18.4% of those 65 or older. In addition, the number is increasing rapidly as seen in Fig. 1, probably due to increasing obesity, decreasing exercise, and a recent change in the definition of diabetes to a lower glucose level. The economic impact is impressive: 1997 estimated direct medical costs of US$ 44 billion and indirect costs of US$ 54 billion for an economic cost annually of about US$ 100

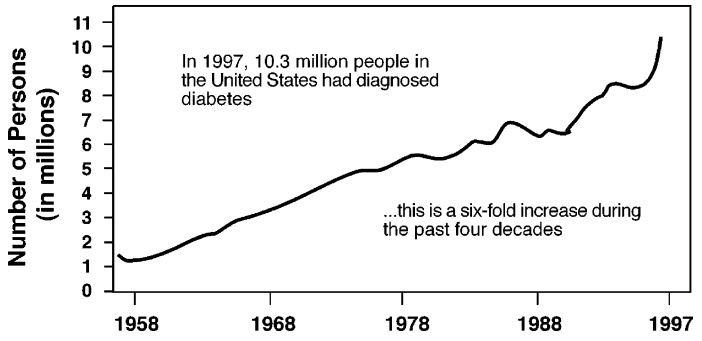

Fig. 1. The rapid growth of diabetes in the United States. Source: American Diabetes Association and the Centers for Disease Control and Prevention.

billion dollars [14]. One in every seven healthcare dollars and 25% of the Medicare budget are spent on diabetic patients [6]. If cost savings can be achieved in diabetes, this can have a significant impact on healthcare spending.

The burden of suffering of diabetes, the seventh leading cause of death in the US, is sadly even more impressive. Death certificates indicate that diabetes contributes to 193,000 deaths annually, but this may be only 36% of the actual count according to one study [36]. Diabetes is the leading cause of new cases of blindness in adults aged 20–74 (as many as 24,000 become blind annually from diabetes), of end-stage kidney disease (33,000 diabetics start dialysis annually), and of leg amputations not related to injury (86,000 annually). Forty percent of new cases of blindness caused by diabetes were in the less than 65 age group in one study [12]. Diabetic patients are 2–4 times more likely to have a heart attack or stroke than a non-diabetic patient. Close to two-thirds of diabetic patients also have hypertension. The social effects are massive with disability among diabetics 2–3 times higher than non-diabetics [31], congenital malformation rates up to 10% in the 18,000 deliveries annually to women with pre-existing diabetes if preconception care is not provided to them, and deaths of newborns at rates 2–3 times higher than average for non-diabetic pregnancies [13].

Diabetes has been well studied, and many of the complications can be prevented. Early detection and proper treatment of diabetes can prevent up to 90% of blindness, and at least 50% of dialysis and amputations [13]. Because of the economic and clinical impact of the disease, a great deal of energy has been put into guidelines, best practices, optimization of care, and other management methods to improve outcomes and save money. In short, possibly all the easily seen low-hanging fruit has been picked already. In this paper, we take some next steps, using data mining, to identify factors that might further reduce the impact of diabetes.

The South has the highest incidence of diabetes per 1000 (35.3) compared with other regions (24.9–26.5) [1, Table 61]. The most recent report indicates that at least 365,000 or 8.4% of Louisiana residents 20 years and older have diabetes, and in Healthcare State Rankings for 2000, Louisiana ranked second worst in the nation in health indicators. Louisiana ranked worst in the nation in diabetes death rate—38.7 deaths per 100,000 population [24]. The State's 1999 data tables show a diabetes death rate of 68.9 for the City of New Orleans [28, Table 26.1]. This is likely due to New Orleans being the fattest city in the United States with 38% of adults being obese [8], and high fat foods beating exercise as a favorite recreation. With this high diabetic prevalence and mortality, it is an especially promising location to investigate a diabetic data warehouse.

3. Understanding the data

The owner of this private diabetic data warehouse is a large integrated healthcare system in the New Orleans area. It combines a 442 bed tertiary care hospital, a 500 physician multi-specialty clinic in 25 locations, one of the largest health plans in the state with 190,000 members, a graduate medical education division with >200 residents/fellows in 25 residency/fellowship programs, and an active research division that includes an outcomes research group.

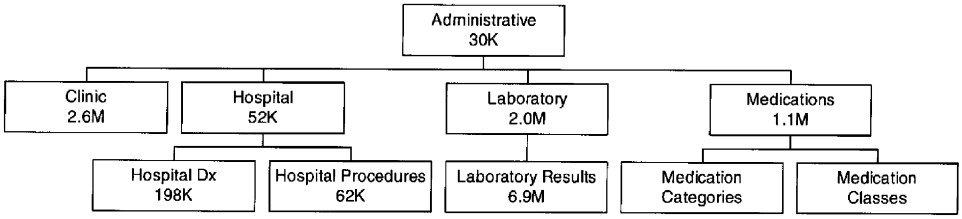

Fig. 2. Diabetic data warehouse structure ($K = 1000$ and $M = 1,000,000$).

The diabetic data warehouse, as of August 2001, included 30,383 diabetic patients. Although the start and termination dates of insurance coverage are not in the data warehouse, the average time in program for active patients was 23.8 months during the 42-month-period from 1 January 1998 through 30 June 2001. This was calculated as the time between the first and last services (clinic, hospital, lab, or medication fill date), and underestimates time in program since many continuity patients may not have had a service performed in the initial or last segments of the study period even though they were active patients.

Fig. 2 outlines the data warehouse structure of the Oracle database along with the number of rows in each table. The database is set up so that the administrative group links the other four groups (clinic, hospital, laboratory, and medication groups). This paper will be limited to aspects of the administrative and clinic databases, and one laboratory test (Table 1).

Understanding the data also involves awareness of its limitations. These data were obtained for purposes other than research. Clinicians will be aware that billing codes are not always precise, accurate, and comprehensive in describing a patient's diagnoses and procedures. However, the codes are widely used in outcomes modeling. Our use of codes in a comorbidity index, later, should be robust. Epidemiologists and clinicians will be aware that important predictors of diabetic outcomes are missing from the database, such as body mass index (BMI), family history of diabetes, time since onset of diabetes, diet and exercise habits. These variables were not electronically stored and would require going to the paper chart and patient interviews to obtain.

BMI might be approximated by an ordinal labeling of patients as 0, 1, or 2 based on whether obesity billing (ICD9) codes are missing, there is an obesity code, or there is a morbid obesity code. However, obesity is not usually a billable diagnosis, so clinicians often do not list it or do so sporadically. This is one of the limitations of a database initially collected for billing and highlights the need to understand not only the data set but how it was collected, why it was collected, and the motivations of those doing the collecting. This will sometimes direct us to avoid some variables as unreliable (e.g. BMI via obesity codes) and consider other variables robust (e.g. comorbidity index, discussed later).

4. Preparation of the data

The importance of the data preparation step cannot be understated, since the value of later data mining or analysis steps depends on it. A key component is data transformation, from the relational structure of the diabetic data warehouse with its multiple tables, to a

Table 1
Selected variables of some subtables within the diabetic data warehouse

Administrative
Clinic_no.: clinic number of patient
Name: name of patient
DOB: date of birth
Sex: sex of patient
Clinic
Year: year of this clinic visit
Month: month of this clinic visit
Clinic_no.: clinic number of patient
POS_code: point of service code describing type of visit
DOS_date: date of clinic visit
Diagnosis_code_1: first of the four ICD9 diagnosis codes for billing
Diagnosis_code_2: second of the four ICD9 diagnosis codes for billing
Diagnosis_code_3: third of the four ICD9 diagnosis codes for billing
Diagnosis_code_4: fourth of the four ICD9 diagnosis codes for billing
Laboratory
Year: year of collection date
Month: month of collection date
Clinic_no.: clinic number of patient
Accession_test_series: a control field used to uniquely identify the record
Collection_date: the collection date
Test_code_ordered: test code that was ordered
Test_code_performed: test code that was performed
Test_name: name of the test
Laboratory Dx (subset of laboratory)
Year: year of collection date
Month: month of collection date
Accession_test_series: a control field used to uniquely identify the record
Result_abbrev.: abbreviation used for the test result
Result: the test result

form suitable for data mining. Data mining algorithms most often are based on a single table, within which there is a record for each individual, and the fields contain variable values specific to the individual. We call this the *data mining data table*. The most portable format for the data mining data table is a flat file, with one line for each individual record. There will be a fixed number of fields (whether each is fixed width or the fields are variable width with a specified separator, is an implementation detail). Some fields may be blank in any given line.

Most often, the data mining data table is constructed by one or more SQL statements on the data warehouse, and the flat file output. The data mining software then reads the flat file. We have taken this approach. Some data mining software has the ability to address the data warehouse directly, thus skipping the intermediate step of a flat file. The steps include:

- Review each table of the relational database and select fields to export.
- Determine the interactions between the tables in the relational database.
- Define the layout of the data mining data table.

- Specify patient inclusion and exclusion criteria, which will determine the number of records in the data mining data table. What is the time interval? What is the minimum and maximum number of records (e.g. clinic visits, or outcome measures) each patient must have to be included? What relevant fields can be missing and still include the individual in the data mining data table?
- Data extraction, including the stripping of patient identifiers.
- Sanity checks on the data mining data table, insuring, e.g. that the minimum and maximum of each variable make clinical sense.

In the data mining data table, the first three fields comprise the administrative variables of Clinic_no., DOB, and sex. We exclude those who do not have continuity, defined as having at least two clinic visits. Any one given patient may have many office visits, but these must all be placed on the same row. Therefore, the variables must be transformed into summary data that contain the most useful information. This avoids having 100 columns for clinic visit 1 through 100 with most rows just being populated in the first few columns. Since each visit has more than a dozen variables associated with it, this actually would mean having >1000 sparsely populated columns for office visits unless summary transformations are used.

Table 2 shows, for each of the three data warehouse tables, Admin, Clinic, Lab (column 1) which fields are exported (column 2). For each exported field or set of fields, the figure lists one or more fields in the data mining data table that are constructed from the exported field or fields (column 3). Multiple—at least two—clinic and laboratory records will contribute to each record in the data mining data table.

While each of these variables needs to be explored in depth, we discuss two in more detail now, the comorbidity index and the average HgbA1c.

The comorbidity variable takes the diagnostic codes in the diabetes registry and converts them into one comorbidity column with one line per patient. We divide all the codes into the 17 categories of the ICD9:

001–139 infectious and parasitic diseases,
140–239 neoplasms,
240–279 endocrine, nutritional and metabolic diseases, and immunity disorders,
280–289 diseases of the blood and blood forming organs,
290–319 mental disorders,
320–389 diseases of the nervous system and sense organs,
390–459 diseases of the circulatory system,
460–519 diseases of the respiratory system,
520–579 diseases of the digestive system,
580–629 diseases of the genitourinary system,
630–679 complications of pregnancy, childbirth, and the puerperium,
680–709 diseases of the skin and subcutaneous tissue,
710–739 diseases of the musculoskeletal system and connective tissue,
740–759 congenital anomalies,
760–779 certain conditions originating in the perinatal period,
780–799 symptoms, signs, and ill-defined conditions,
800–999 injury and poisoning.

Table 2
Transformed variables

Variable	Data warehouse field	Data mining data table field	Notes
Admin	Clinic_no.	ID	Patient identifier is removed, ID is simply a record number 1, 2,... in the data mining data table
	DOB	Age	Transform DOB to age in years as of 1 January 2001
	Sex	Sex	Exclude those listed as ''U'' (unknown) or ''?''—this was only 6 of >30,000
Clinic	Clinic_no.		For record linking
	POS_code (point of service)	ER	Number of ER visits in the given time period
		OV	Number of office visits in the given time period
	DOS_code (date of service)	TBD	The specific dates within the time period of interest when the ER or office visits occurred. There is sequencing and time-series information, but data mining software will not be able to utilize this without summary transformations. Use for this field is to be determined.
	Diagnosis_code_1–4	Comorbidity	Number of major body systems (1–17) listed as a diagnosis code. This can act as a rough comorbidity index. See details in text
		Lipids	Is there a lipid disorder documented
		HTN	Is there a hypertensive disorder documented
		CV	Is there a CAD/PVD disorder documented
		Eye	Is there retinopathy documented
		ESRD	Is there ESRD/CRF/CRI documented
Lab	Assession_test_series		For record linking
	HgbA1c (may have many different values from blood tests on various dates)	Av_HgbA1c	Summary measure of the average hemoglobin A1c. A binary variable (0, 1) is created using a cut-point of 9.5 in the average HgbA1c

We then label patients as having 1 through 17 of the categories they have been diagnosed into as a rough comorbidity index. For example, if someone had codes through their visits that fell into five of these groups, their comorbidity index would be 5. Many comorbidity indexes, such as Charlson's index, are inpatient focused, whereas most of the diabetic patients in our study were never hospitalized. Thus, we agree with others who have found it valuable to use a physicians' claims comorbidity index [27]. This is a simplified variant of that idea.

Handling time-series medical data is challenging for data mining software. One example in our study is the HgbA1c value, the key measure of glycemic control that should be measured every 3–6 months in all diabetics. This is closely related to clinical outcomes and complication rates in diabetes. Healthcare costs increase markedly with each 1% increase in baseline HgbA1c; patients with an HgbA1c of 10% versus 6% had a 36% increase in 3 years medical costs [6]. How should this time-series variable be transformed from the relational database to a vector (column) in the data mining data table? A given diabetic patient may have many of these HgbA1c results. We could pick the last one, the first, or a median value. We could take an average. Since the trend over time for this variable is important, we could choose the slope of its regression line over time. However, a linear function may be a good representation for some patients, but a very bad one for others that may, for example, be better represented by an upside down U-curve. This difficulty is a problem for most repeated laboratory tests. In any event, some information will be lost.

In this study, we chose to use the average HgbA1c of all the results for a given patient and excluded patients who do not have at least two HgbA1c results in the data warehouse. As noted in the previous table, we repartitioned this average HgbA1c into a two-level categorical variable based on a meaningful clinical cut-point of 9.5. Experts agree that an HgbA1c >9.5 is a bad outcome, or a medical quality error, no matter what the circumstances [2].

Our final data mining data table had 15,902 patients (rows) and included the 11 variables listed in Table 2 above. All of these patients had at least two HgbA1c tests and at least two office visits, the criteria we used for minimal continuity in this 3.5-year-period.

5. Data mining

There are many data mining methods. Here, we used the classification tree approach as standardized in the CART software by Salford Systems. As detailed in [22], the principle behind all tree models is to recursively partition the input variable space to maximize purity in the terminal tree nodes. The partitioning split in any cell is done by searching each possible threshold for each variable to find the threshold split that leads to the greatest improvement in the purity score of the resultant nodes. This is recursively done for each additional node. Hence, this is a monothetic process, that is, each node is split on just one variable, which may be a limitation of this method in some circumstances.

In CART's defaults, the Gini splitting criterion is used, although one can opt for a number of other methods. Although, this could recursively continue to the point of perfect purity which would sometimes mean only one patient in a terminal node, this overfitting of the data does not help in accurately classifying another data set. Therefore, we divide the

data randomly into learning and test sets. The number of trees generated is halted or pruned back by how accurately the classification tree created from the learning set can predict classification in the test set. Cross validation is another option for doing this, though in the CART software's defaults this is limited to $n = 3000$. This could be changed higher to use our full data set, but some CART consultants note "The n-fold cross-validation technique is designed to get the most out of datasets that are too small to accommodate a hold-out or test sample. Once you have 3000 records or more, we recommend that a separate test set be used" [34]. The original CART creators recommended dividing the data into test and learning samples whenever there were more than 1000 cases, with cross validation being preferable in smaller datasets [11].

The 10 predictor variables were used with the binary target variable of the HgbA1c average (cut-point of 9.5) in an attempt to find interesting patterns that may have management or clinical importance and are not already well known. CART 4.0 gives the tree in Fig. 3, when the target variable is HgbA1c >9.5 (0, 1), the predictors are age, sex, ER, OV, CMI, lipid, HTN, CV, eye, ESRD, and the software defaults are used except for testing via a random variable to separate learn and test samples.

The variables that are most important to classification in the optimal CART tree are listed in Fig. 4, which is reproduced from the actual CART output. CART can be used for multiple purposes. Here, we want to find clusters of deviance from glycemic control. With no analysis, there is a 1052/7953 or a 13.2% rate of bad gylcemic control in the learning sample. This is the same as odds of 1 in 6.56. We want to find nodes that have higher rates or odds of bad glycemic control. The very first split in node 1 of the above tree is based on an age of 65.6 (CART sends all cases less than or equal to the cut-point to the left and greater than the cut-point to the right). In node 2 ($\leq$65.6 years of age) we have bad glycemic control in 775/3987 or 19.4% or odds of 1 in 4.14.

If we look at the tree to find terminal nodes (TN) where the percentage of all the patients in those nodes have a bad HgbA1c that is even greater than the 19.4% true of those $\leq$65.6 years of age, we identify the TNs listed in Table 3 in the learning sample.

The purer the nodes we limit ourselves to, the less percentage of the overall population with bad glycemic control we get. With no analysis in the learning set, we can capture all 1052 patients with bad glycemic control but must target the entire group of 7953. When we

Table 3
Terminal nodes with percentages of bad HgbA1c values greater than the first split with age less than 65.6

TN no.	w/bad HgbA1c (%)	Number w/bad HgbA1c	N	Rules
1	23.8	519	2185	age $\leq$55.231
4	30.8	8	26	(age $\leq$57.781) and (comorbidity >6.5) and (ER >3.5) and (lipid = 1)
6	33.3	13	39	(age >55.231) and (comorbidity >6.5) and (ER >3.5) and (lipid = 0)
8	22.5	23	102	(65.581 < age $\geq$ 72.788) and (OV $\leq$47.5) and (ER $\leq$4.5) and (lipid = 0) and (comorbidity $\leq$3.5)
Total	23.9	563	2352	

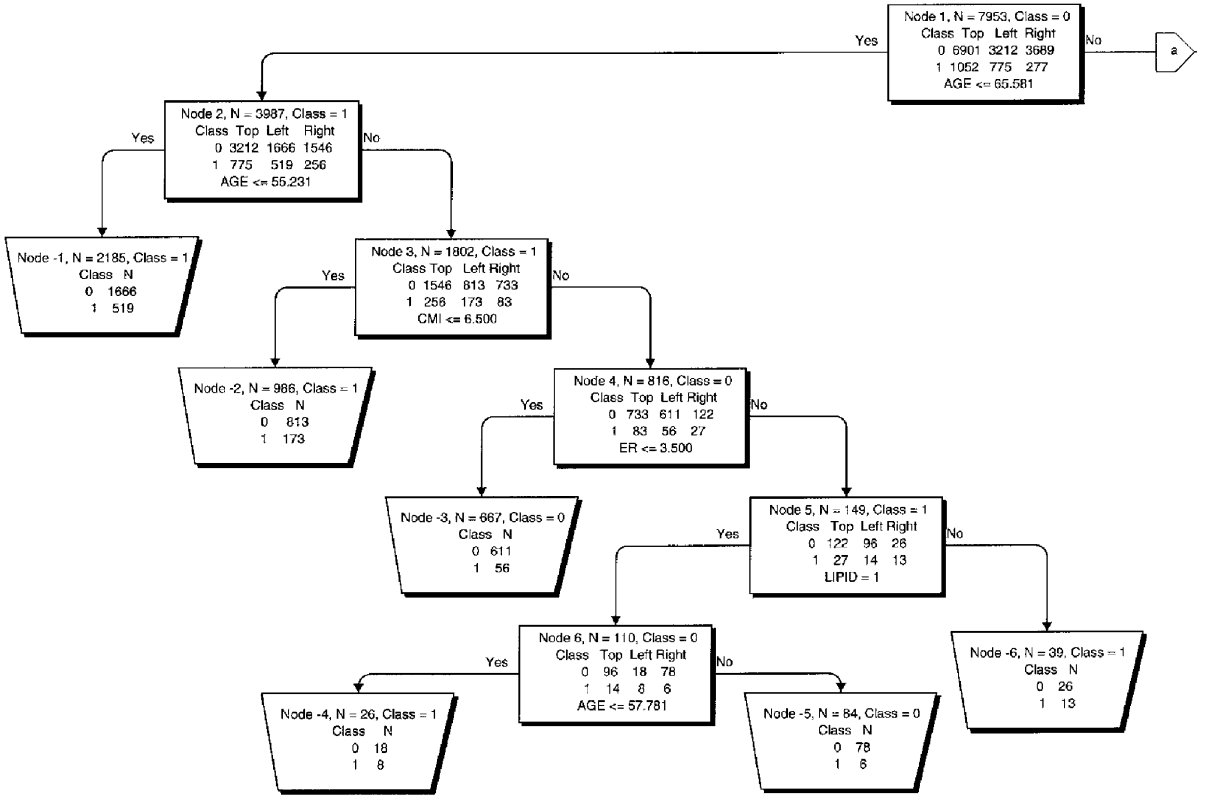

Fig. 3. CART tree output using defaults with target variable of HgbA1c >9.5 (yes = 1, no = 0). Each decision node (1–12) lists the number of each target class (0, 1) that goes to the left if the split criterion is met and to the right if it is not met. Each terminal node (−1 to −13) lists the number of each target class and what the node is classified as.

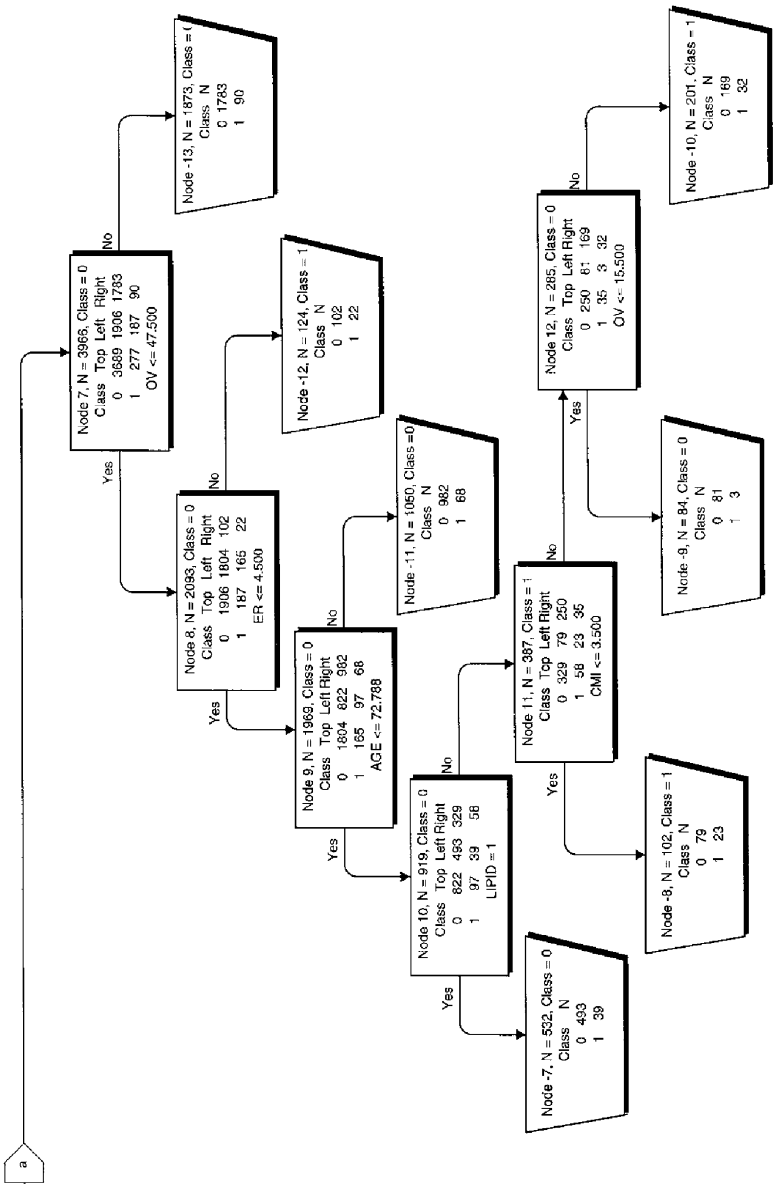

Fig. 3. (*Continued*).

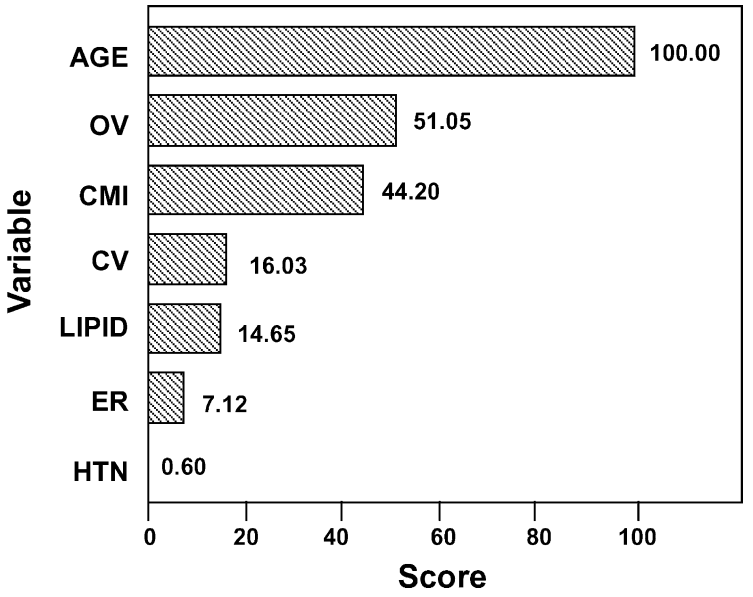

Fig. 4. Variables most important to classification of target variable.

limit ourselves to the purer node 2 in the above tree, we capture 775 or 74% of those with bad glycemic control by targeting only 3987 or 50% of the population who are <65.6 years of age. If we limit ourselves to TN1, we capture 519 or 49% of those with bad glycemic control by targeting only 2185 or 27% of the population who are <55.2 years of age. If we use more complicated criteria by combining TNs 1, 4, 6, and 8 we capture 563 or 54% of those with bad glycemic control by targeting only 2352 or 30% of the population. There are issues of resources and cost-effectiveness for managers to sort out in deciding upon intervention targets.

The classification error in the learning and test samples are substantial as evidenced in Tables 4–6 where a quarter of the bad glycemic control patients are missed in the CART analysis. CART is doing a good job with the 10 predictor variables it is given, but more accurate prediction requires additional variables. Our future work will include additional hospital, lab, and pharmaceutical variables that should decrease misclassification. This particular electronic record does not contain other key variables such as BMI mentioned in Section 3 and this will limit how low the misclassification rate can be pushed down even with many other variables.

Table 4
Classification error in learning sample

Learning sample			
Class	*N* cases	*N* mis-classed	Pct error
1	1052	262	24.90
0	6901	2873	41.63

Table 5
Classification error in test sample

Learning sample			
Class	*N* cases	*N* mis-classed	Predicted error
1	1060	303	28.58
0	6890	2919	42.37

Table 6
CART analysis test sample classification table

Actual class	Predicted class		Actual total
	0	1	
0	3971.00	2919.00	6890.00
1	303.00	757.00	1060.00
Predicted total	4274.00	3676.00	7950.00
Correct	0.576	0.714	
Success individual	−0.290	0.581	
Total correct	0.595		

Sensitivity: 0.576; specificity: 0.714; false "0": 0.071; false "1": 0.794.

Adjustment to defaults in CART can give better results. For example, if the splitting rule is expanded to include linear combinations with Gini, we get the smaller tree in Fig. 5. Note that the formula for splitting at the parent node is now 0.451 (age) + 0.893 (CMI) $\leq$ 32.5576. This gives us TN1 where we capture 667 or 63% of those with bad glycemic control by targeting only 3011 or 38% of the population. While this is mathematically attractive compared with some of the above choices, this is a more complicated population for managers to target. Not only must managers identify the persons (easy enough with

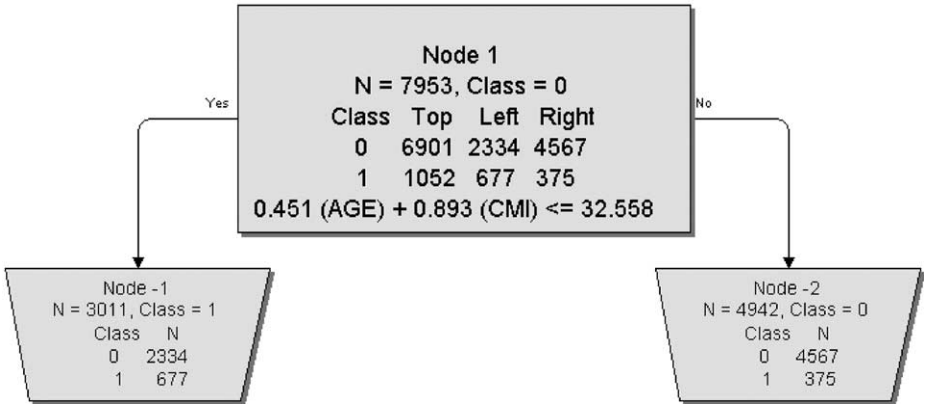

Fig. 5. CART tree output using Gini with linear combinations.

computers), but they must get enough of a feel for what the population characteristics are to know what interventions are likely to be helpful. This is more intuitive for those who are <55 or <65 than it is for those who satisfy 0.451 (age) + 0.893 (CMI) $\leq$ 32.5576.

6. Evaluation of the discovered knowledge

From this CART analysis, there is a clinically surprising observation. The most important variable associated with a bad HgbA1c score is younger age, not the comorbidity index or whether patients have related diseases. The first level of the tree shows that just dividing people using an age cut-point of 65.581 years of age, 19.4% of younger people ($n = 3987$) have a bad HgbA1c. This is 2.8 times the rate of bad HgbA1c values in those who are older (7.0%, $n = 3966$). In other words, those less than 65.6 years of age are almost three times as likely to have bad glycemic control than those who are older. This is surprising information to most clinicians. The age information for the learning group is shown in Table 7.

We see that those with bad HgbA1c values are 1052/7953 or 13.2% of all diabetic patients in the learning group. The odds of having a bad HgbA1c is 1052/6901 or 0.152. If we calculate the odds of a bad HgbA1c in the younger and older groups, we can then calculate an odds ratio (OR). Going to the original data on all 15,903 patients, we have the odds of a bad HgbA1c in the younger group ($\leq$65.581 years of age, $n = 8000$) is 1555/6445 or 0.241. The odds of a bad HgbA1c in the older group ($n = 7903$) is 557/7346 or 0.0758. Therefore, the odds ratio that someone is less than 65.6 years of age if they have a bad HgbA1c (average reading >9.5) is 0.241/0.0758 or 3.18 [18]. The 95% confidence interval for this OR is (2.87, 3.53) using EpiInfo2000 (epidemiologic software that can be downloaded from the Centers for Disease Control and Prevention at www.cdc.gov/epiinfo/index.htm).

Similarly, the odds ratio that someone is less than 55.2 years of age rather than >65.6-year-old if they have a bad glycemic control is (519/1666)/(557/7346) or 4.11 (3.60, 4.69)!

If we want to target diabetics with bad HgbA1c values, the odds of finding them are 3.2 times as high in diabetic patients <65.6 years of age than those older and 4.1 times as high in those <55 than over 65. We can use the TN information in the earlier section to identify the highest risk groups for having a bad HgbA1c, though the vast majority are simply diabetic patients who are less than 55.2 years of age. This is clinically important information because the younger group has so many more years of life left to develop diabetic complications from bad glycemic control. This is especially helpful because this

Table 7
Learning group ages (mean, standard deviation) for HgbA1c values

Age	Group		
	HgbA1c $\leq$9.5	HgbA1c >9.5	All HgbA1c
Mean	64.759	56.576	63.676
S.D.	13.272	13.711	13.615
N	6901	1052	7953

tells us which population to target interventions at even before we have the HgbA1c values to show us.

7. Using the discovered knowledge

It appears this surprising information, that younger age significantly predicts bad diabetic control, will be of assistance to clinic management in targeting more focused interventions. Physicians can be alerted to the high probability that younger diabetic patients will have poor diabetic control, and more organized follow-up of their glycemic control may be needed. Health maintenance organizations and public health workers may want to explore what educational interventions can be successfully directed to younger diabetics (<65, especially <55) who are 3.2 or 4.1 times as likely to have bad glycemic control than the geriatric diabetic patients.

Similar diabetic data mining studies in other geographic and cultural locations are needed to see if younger age significantly predicts bad diabetic control beyond the New Orleans area. If this were confirmed with multiple regional or national diabetic data warehouses, some of the interventions discussed earlier would have national significance well beyond the clinic that owns the data warehouse on which the initial study was performed.

We are actively data mining this diabetic data warehouse and expect to find additional useful information as we expand predictors to include multiple laboratory, hospital, and pharmaceutical variables. Target variables will also be expanded to include hospitalizations, costs, percentage of medications filled, and key clinical outcomes such as myocardial infarctions or cerebrovascular accidents.

Management will need to balance the cost effectiveness of targeting particular populations to improve outcomes with the realistic hope of success. For example, a simple cut-point of younger age is something that public health and clinical workers could easily use in developing useful interventions for this group. A complicated formula to identify those with poor glycemic control, even if mathematically more accurate, may not be helpful in knowing how to develop useful interventions in that population beyond what you would do for everyone.

8. Conclusions

Data mining is valuable in discovering novel associations that can prove useful to clinicians and administrators as noted earlier. However, areas that need further work to fully utilize data mining in healthcare include time-series issues, sequencing information, data squashing technologies, and a tight integration of domain expertise and data mining skills.

We have already discussed time-series issues. This has been investigated [3–5,7,20, 26,29,30,35], but there is a great need to explore this further in healthcare data mining.

The sequence of various events may hold meaning important to a study. For example, a patient may have better glycemic control, manifested in improved HgbA1c values, especially when the patient had an office visit with a physician within a month of a

previous HgbA1c. Perhaps this is meaningful information that implies that the proper sequence of physician visits relative to HgbA1c measurements is an important predictor of good outcomes. All this information is located in the relational database, but we must ferret it out by having in advance an idea that a sequence of this sort may be important and then searching for such associations. There may be many such sequences, involving interactions between hospital, clinic, pharmacy, and lab variables. Regrettably, no amount of data mining will be able to extract sequence associations where we have not thought to extract the prerequisite variables from the relational database into the data mining data table. In the ideal data mining scenario, software could interface directly with the relational database and extract all possibly meaningful sequences for us to review, and domain experts would then sort through the list. This issue has begun to get attention [17] and will need to be addressed in future healthcare data mining.

It has been shown that a data squashing algorithm to reduce a massive data set is more powerful and accurate than using a random sample [16]. Squashing is a form of lossy compression that attempts to preserve statistical information [15]. In our study, we used all the data available in the CART analysis and did not sample. If a massive dataset were used, such as all diabetics in a nationwide Medicare database, memory and time constraints may require limitations on numbers of observations used. The newer data squashing techniques may be a better approach than random sampling in these massive datasets.

Transactional healthcare data mining, exemplified in the diabetic data warehouses discussed above, involves a number of tricky data transformations that require close collaboration between domain experts and data miners [10]. Even with ideal collaboration or overlapping expertise, we need to develop new ways to extract variables from relational databases containing time series and sequencing information. Part of the answer lies in collaborative groups that can have additional insights. Part of the answer lies in the further development of data mining tools that act directly on a relational database without transformation to explicit data arrays.

Acknowledgements

Leonard Medal helped with SQL statements. The CART software was funded by a Grant from GlaxoSmithKline Pharmaceuticals. The Institutional Review Board at the institution that owns the diabetic data warehouse approved this study.

References

[1] Adams PF, Hendershot GE, Marano MA. National Health Interview Survey (US) and National Center for Health Statistics (US). Current Estimates from the National Health Interview Survey, 1996. US Department of Health and Human Services. Centers for Disease Control and Prevention. Hyattsville (MD): National Center for Health Statistics, 1999.

[2] AMA, JCAHO and NCQA, Co-ordinated performance measurement for the management of adult diabetes: a consensus statement from the American Medical Association. The joint commission on accreditation of healthcare organizations, and the National Committee for Quality Assurance, 2001. http://www.ama-assn.org/ama/upload/mm/370/diabetes.pdf.

[3] Bellazzi R, Larizza C, Magni P, Montani S, Stefanelli M. Intelligent analysis of clinical time series: an application in the diabetes mellitus domain. Artif Intell Med 2000;20:37–57.
[4] Bellazzi R, Magni P, Larizza C, De Nicolao G, Riva A, Stefanelli M. Mining biomedical time series by combining structural analysis and temporal abstractions. Proc AMIA Symp 1998:160–4.
[5] Bjorvand AT. Time series and rough sets. Trondheim (Norway): Department of Computer Systems and Telematics, University of Trondheim, 1996. p. 75.
[6] Blonde L. Epidemiology, costs, consequences, and pathophysiology of type-2 diabetes: an American epidemic. The Ochsner J 2001;3:126–31.
[7] Blum RL. Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical database: the RX project. Comput Biomed Res 1982;15:165–87.
[8] Bragg R. A city of high calories and people proud of it. New York: The New York Times, 1997. p. v146 (March 15) p6(N) p8(L) column 1.
[9] Breault JL. Data mining diabetic databases: are rough sets a useful addition? In: Goodman A, Smyth P, Ge X, Wegman E, editors. Proceedings of the 33rd Symposium on the Interface, Computing Science and Statistics, Fairfax, VA. Costa Mesa (CA): The Interface Foundation of North America, in press.
[10] Breault JL, Goodall CR. Mathematical challenges of variable transformations in data mining diabetic data warehouses. In: Proceedings of the Conference on Mathematical Challenges in Scientific Data Mining, 2002 Jan 14–18. Los Angeles (CA): UCLA Institute of Pure and Applied Mathematics, 2002.
[11] Breiman L. Classification and regression trees. Belmont (CA): Wadsworth, 1984. p. 307.
[12] CDC. Blindness caused by diabetes, 1987–1994, MA. MMWR Morb Mortal Wkly Rep 1996; 45:937–41.
[13] CDC. Diabetes: a serious public health problem, 2001. http://www.cdc.gov/diabetes/pubs/glance.htm.
[14] CDC. National Diabetes Fact Sheet, 1998. http://www.cdc.gov/diabetes/pubs/facts98.htm.
[15] DuMouchel W. Data squashing: constructing summary data sets. In: Goodman A, Smyth P, Ge X, Wegman E, editors. Proceedings of the 33rd Symposium on the Interface, Computing Science and Statistics, Fairfax, VA. Costa Mesa (CA): The Interface Foundation of North America, in press.
[16] DuMouchel W, Volinsky C, Johnson T, Cortes C, Pregibon D. Squashing flat files flatter. In: Chaudhuri S, Madigan D, editors. KDD-99. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999 Aug 15–18, San Diego, CA. New York: Association for Computing Machinery, 1999. p. 6–15.
[17] Dzeroski S, Lavrac N. Relational data mining. Berlin: Springer, 2001.
[18] Fos PJ, Fine DJ. Designing health care for populations: applied epidemiology in health care administration. San Francisco: Jossey-Bass, 2000.
[19] Goodall C. Massive data sets in healthcare. In: Proceedings of the Massive Data Sets Committee on Applied and Theoretical Statistics. National Academy of Sciences. Washington, DC: National Research Council, 1995. Online publication at http://bob.nap.edu/html/massdata/media/cgoodall-t.html.
[20] Goodall CR. Data mining of massive datasets in healthcare. J Comput Graph Stat 1999;8:620–34.
[21] Gunopulos D, Das G. Time-series similarity measures (tutorial PM-2). In: Ramakrishnan R, Stolfo S, editors. KDD-2000: Tutorials. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000 Aug 20–23, Boston, MA. New York: Association for Computing Machinery, 2000. p. 243–307.
[22] Hand DJ, Mannila H, Smyth P. Principles of data mining. Cambridge (MA): MIT Press, 2001. p. 343–7
[23] He H, Koesmarno H, Van T, Huang Z. Data mining in disease management: a diabetes case study. In: Mizoguchi R, Slaney JK, editors. PRICAI 2000. Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence, 2000 Aug 28–Sept 1, Melbourne, Australia. Topics in artificial intelligence. Berlin: Springer, 2000. p. 799.
[24] Hood D. Louisiana health report card, 2001. State of Louisiana. Baton Rouge (LA): Department of Health and Hospitals. http://www.dhh.state.la.us/OPH/statctr/4Report%20Card/2001/2001LouisianaHealth-ReportCard.pdf.
[25] Hsu W, Lee ML, Liu B, Ling TW. Exploration mining in diabetic patients databases: findings and conclusions. In: Ramakrishnan R, Stolfo S, editors. KDD-2000. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000 Aug 20–23, Boston, MA. New York: Association for Computing Machinery, 2000. p. 430–6.

[26] Huang YW, Yu PS. Adaptive query processing for time-series data. In: Chaudhuri S, Madigan D, editors. KDD-99. Proceedings Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999 Aug 15–18, San Diego, CA. New York: Association for Computing Machinery, 1999. p. 282–6.
[27] Klabunde CN, Potosky AL, Legler JM, Warren JL. Development of a comorbidity index using physician claims data. J Clin Epidemiol 2000;53:1258–67.
[28] LADHH, 1999. Data Tables, 2000. Louisiana state center for health statistics, Department of Health and Hospitals, Office of Public Health. http://www.dhh.state.la.us/OPH/statctr/1Tables/1999/Parish/t26_99i.xls.
[29] Riva A, Bellazzi R. Intelligent analysis techniques for diabetes data time series. In: Lasker GE, Liu X, editors. Advances in Intelligent Data Analysis. Baden−Baden (Germany): IIAS Press, 1995. p. 144–8.
[30] Sakamoto N. Object-oriented development of a concept learning system for time-centered clinical data. J Med Syst 1996;20:183–96.
[31] Songer TJ. Disability in diabetes. In: National Diabetes Data Group (US), National Institute of Diabetes and Digestive and Kidney Diseases (US) and National Institutes of Health (US), editors. Proceedings of the conference on Diabetes in America. 2nd ed. Bethesda (MD): National Institutes of Health National Institute of Diabetes and Digestive and Kidney Diseases, 1995. p. 259–82.
[32] Stepaniuk J, Rough set data mining of diabetes data. In: Ras Z, Skowron A, editors. Proceedings of the 11th International Symposium on Foundations of Intelligent Systems, 1999 June 8–11 Warsaw, Poland. ISMIS 1999. Berlin: Springer, 1999. p. 457−65.
[33] Tafeit E, Moller R, Sudi K, Reibnegger G. ROC and CART analysis of subcutaneous adipose tissue topography (SAT-Top) in type-2 diabetic women and healthy females. Am J Hum Biol 2000;12:388–94.
[34] Timberlake Consultants. CART frequently asked questions, 2001. http://www.timberlake.co.uk/software/cart/cartfaq1.htm#q23.
[35] Tsien CL. Event discovery in medical time-series data. Proc AMIA Symp 2000:858–62.
[36] Weng C, Coppini DV, Sonksen PH. Linking a hospital diabetes database and the national health service central register: a way to establish accurate mortality and movement data. Diabet Med 1997;14:877–83.