Developing a Prediction Rule From Automated Clinical Databases to Identify High-Risk Patients in a Large Population With Diabetes

JOE V. SELBY, MD, MPH
ANDREW J. KARTER, PHD
LYNN M. ACKERSON, PHD
ASSIAMIRA FERRARA, MD
JENNIFER LIU, MPH

OBJECTIVE — To develop and validate a prediction rule for identifying diabetic patients at high short-term risk of complications using automated data in a large managed care organization.

RESEARCH DESIGN AND METHODS — Retrospective cohort analyses were performed in 57,722 diabetic members of Kaiser Permanente, Northern California, aged ≥19 years. Data from 1994 to 1995 were used to model risk for macro- and microvascular complications ($n = 3,977$), infectious complications ($n = 1,580$), and metabolic complications ($n = 316$) during 1996. Candidate predictors ($n = 36$) included prior inpatient and outpatient diagnoses, laboratory records, pharmacy records, utilization records, and survey data. Using split-sample validation, the risk scores derived from logistic regression models in half of the population were evaluated in the second half. Sensitivity, positive predictive value, and receiver operating characteristics curves were used to compare scores obtained from full models to those derived using simpler approaches.

RESULTS — History of prior complications or related outpatient diagnoses were the strongest predictors in each complications set. For patients without previous events, treatment with insulin alone, serum creatinine ≥1.3 mg/dl, use of two or more antihypertensive medications, HbA_{1c} >10%, and albuminuria/microalbuminuria were independent predictors of two or all three complications. Several risk scores derived from multivariate models were more efficient than simply targeting patients with elevated HbA_{1c} levels for identifying high-risk patients.

CONCLUSIONS — Simple prediction rules based on automated clinical data are useful in planning care management for populations with diabetes.

Diabetes Care 24:1547–1555, 2001

Approximately 4% of most managed care populations have diabetes (1,2), but these patients account for nearly 12% of total health care expenditures (2). These costs also reflect catastrophic events in the lives of patients, because a large fraction of total costs result from hospitalization for disease complications (2). Recognizing the burden of this illness, many managed care organizations have developed intensive diabetes care programs, featuring multidisciplinary clinics or nurse case management (3–5) to improve pharmacotherapy, preventive screening, and support for self-care. However, inclusion of all diabetic patients in intensive disease management programs, including those at low risk for complications, would diminish the cost-effectiveness of these programs.

Clinical prediction rules (6) are tools created by combining information from clinical data, usually using multivariate analyses, to estimate the probability of an outcome for individual patients. When applied to an entire population of members with diabetes, a prediction rule could be used to identify and rank members by their level of risk for complications. Despite the frequent availability of rich automated clinical data in health plan systems (7), prediction rules have not been widely used for diabetes. Instead, many programs focus solely on poor control of HbA_{1c} levels to identify those in need of more intensive intervention. This study seeks to develop and test a prediction rule using automated clinical data that can be applied at the population level to improve this strategy. Several approaches are compared to identify the simplest rule that efficiently identifies high-risk patients.

RESEARCH DESIGN AND METHODS

— This report is based on a retrospective cohort analysis conducted in the Northern California Kaiser Permanente Diabetes Registry. Kaiser Permanente, which is a group model health maintenance organization (HMO), had ~2.5 million enrollees during the study period. The registry (2) is an ongoing epidemiological cohort of all HMO members with diabetes identified from four automated databases: pharmacy prescriptions for diabetes medications, abnormal HbA_{1c} values (≥6.7%) in laboratory files, primary hospital discharge diagnoses of diabetes, and emergency department records of diabetes as the reason for visit. During the study period, this registry had a sensitivity of 90% when matched against >1,500 self-reported diabetic patients who responded to two large mailed

From the Division of Research, Kaiser Permanente of Northern California, Oakland, California.
Address correspondence and reprint requests to Joe V. Selby, Division of Research, Kaiser Permanente, 3505 Broadway, Oakland, CA 94611. E-mail: jvs@dor.kaiser.org.

Received for publication 9 January 2001 and accepted in revised form 12 April 2001.

Abbreviations: AUC, area under the curve; HMO, health maintenance organization; OR, odds ratio; ROC, receiver operating characteristics.

A table elsewhere in this issue shows conventional and Système International (SI) units and conversion factors for many substances.

surveys. The registry missed some diet-controlled subjects and the small proportion of members who never use the HMO's services. The registry also has been found to contain ~2.5% false positives or members who do not truly have diabetes (J.V.S., unpublished data).

For these analyses, data gathered from electronic databases and from a mailed survey during the 2-year baseline period (1994–1995) were used to predict three sets of complications of diabetes occurring during 1996. The study sample consisted of the 57,722 registry members who were aged ≥19 years and who were known to have diabetes by 1 January 1994; they were continuously enrolled in the health plan throughout the 2-year baseline period (1 January 1994 to 31 December 1995) and remained in the health plan for at least the first month of 1996. Continuous enrollment was defined as not having a membership gap of >2 months' duration. We excluded 970 patients with no outpatient utilization (visits, laboratory tests, or prescriptions) during the baseline period, because they would provide few data on predictors, and they may well have received care, including care for complications, outside the HMO.

Study outcomes

Outcomes were complications requiring hospitalization during 1996; they were identified from principal discharge diagnoses in the HMO's discharge databases (for 16 plan hospitals and for all claims received from out-of-plan hospitals). Complications were grouped into three sets: macro- and microvascular, infectious, and metabolic; they are listed by *International Classification of Diseases, 9th Revision, Clinical Modification* discharge diagnosis code in Table A1 in the APPENDIX. Dichotomous dependent variables were created to indicate whether one or more complications from each set were noted during 1996.

Candidate predictors

The 36 potential predictors of complications (during the baseline period) are shown in Table A2 in the APPENDIX. These included hospital discharges for the same complications sets as the study outcomes and the outpatient diagnoses that are related to these complications. Dichotomous predictor variables were used to note occurrence during the baseline period of any complication-related hospitalization. Dichotomous predictors were also used to indicate the baseline presence of outpatient diagnoses that are closely related to either macro- and microvascular or infectious complications. For example, renal insufficiency and unstable angina are likely to be important predictors of future hospitalizations. No outpatient diagnoses related to metabolic complications were captured in this HMO's data systems.

Other candidate predictors included laboratory results (HbA_{1c}, serum creatinine, and lipoprotein levels), pharmacy prescriptions (for hypoglycemic, lipid-lowering, and antihypertensive agents), outpatient visit counts by type, and responses to a 1994–1996 mailed survey (with telephone follow-up of nonrespondents) completed by 83% of the study sample. Survey items included demographics, self-reported behaviors, and information used to classify diabetes (age and obesity status at onset and patterns of insulin use).

Both clinical and survey databases have relatively high rates of missing data for potential predictors. Approximately 25–40% of cohort members had missing values for one or more key predictors, such as baseline HbA_{1c}, serum creatinine, cholesterol, smoking status, and BMI. Because we wished to develop a tool applicable to an entire population, it was important that these subjects were included in the models. We therefore used "missing" categories for several key variables. Continuous predictors were converted to ordered categorical variables for this purpose.

Data analyses

The cohort of 57,722 members was randomly split into derivation and validation data sets. Of all the subjects, >94% remained under observation throughout 1996. Of the remaining 6%, 52% were censored because of death rather than leaving the health plan. Because of nearly complete follow-up and a short observation period, we used logistic regression to model the data.

After examining bivariate associations of predictors and outcomes, separate stepwise logistic regression models were conducted in the derivation data set to build the best model for each complications set. On parameter estimates, $P < 0.01$ was required to include a predictor in each best model. Once each model was derived, coefficients for significant predictors were applied to predictor values of the validation data set members. Risk scores for each member were calculated by summing coefficients across all predictors, and the ability of these scores to predict complications in a new population was examined.

Based on preliminary analyses, four simpler approaches to identifying and targeting high-risk patients were identified and compared with the best model. At an early stage in our analyses, we noted that events or related outpatient diagnoses during the baseline period were strong predictors of each complications set. Therefore, the first alternative was to use a "prior events" strategy that simply targeted patients with either of these predictors. Preliminary analyses also revealed that risk scores based only on the first three variables entering each model were nearly as sensitive as scores from the best models. Therefore, we evaluated "reduced models" that included only these first three variables.

The third comparison approach tested a simplified numerical risk score derived by replacing significant model coefficients with integer values as follows: a value of 1.0 for a (significant) multivariate odds ratio (OR) between 1.1 and 1.49, 2.0 for an OR between 1.50 and 1.99, and 3.0 for an OR of ≥2.0, with corresponding negative numbers for significant ORs <1.0. To obtain integer values for age, which was the only continuous variable in any model, we calculated the age-specific OR distribution (relative to 20 years of age, which was the youngest age possible) using the model coefficients for 10-year increases in age and applied the same OR cut points to categorize the distribution into values from 0 to 3. The integer values were summed to yield a simple numerical score. If this approach performs nearly as well as the risk score from the best model, it yields a much simpler algorithm for use in other populations.

The fourth strategy was to simply rank patients on the basis of their average HbA_{1c} level during 1994–1995 and to select patients in descending order of these values. We used percentiles rather than absolute values for cut points because HbA_{1c} distributions vary across populations and over time.

Initial comparisons of these five approaches focused on sensitivity and positive predictive values in the validation data set. Continuous risk scores, which identified the 30% of patients with the highest predicted risk (or the highest HbA_{1c} levels), were compared at the cut point. This cut point was chosen to be consistent with our health plan's current policy of planning more intensive interventions for ~30% of the population. Given its distribution, the numerical score, which is ordinal, was cut as close to the upper 30th percentile as possible. For the prior events approach, the proportion with such an event is fixed. Continuous and ordinal scores were also compared across their entire range of values. Differences in areas under the curve (AUCs) of receiver operating characteristics (ROC) curves were tested with ROC Analyzer (8,9), which uses a nonparametric method of estimating AUC and adjusts for the correlation of the two curves (10).

For patients without prior inpatient events or related outpatient diagnoses, we re-examined the utility of the four remaining approaches. In this subgroup, the number of macro- and microvascular complications, infectious complications, and metabolic complications was greatly reduced, leaving just 723 subjects who experienced at least one event in 1996 (561 with a macro- and microvascular event, 453 with an infectious event, and 95 with a metabolic event). Because all complications are important from a disease management perspective, and in light of the overlap of many important predictors for two or all three sets of complications, we combined these end points and modeled risk for any event. Age, sex, and race were not included in this model, despite associations with one or more outcomes in the models described above, because these characteristics present no options for risk reduction. By removing them from the models, many of the associated, mutable risk factors should contribute more strongly to risk scores. We further excluded the 3.5% of remaining patients with serum creatinine levels ≥2.0, reasoning that these subjects should already be targeted because of their known and very high-risk status.

RESULTS — Number of subjects, demographic characteristics, and frequency of each set of complications were similar in the derivation and validation data sets (Table 1). Macro- and microvascular events occurred nearly three times as frequently as infectious events and >10 times as frequently as metabolic complications.

Table 1—*Demographic and clinical characteristics of diabetic patients in derivation and validation data sets*

	Derivation data set	Validation data set
n	28,838	28,884
Mean years of age (range)	60.8 (19–101)	60.5 (19–99)
Percent female	46.7	47.6
Race [*n* (%)]		
African-American	3,513 (12.2)	3,611 (12.5)
Asian/Pacific Islander	3,080 (10.7)	3,054 (10.6)
Hispanic	3,537 (12.3)	3,600 (12.5)
White	15,292 (53.0)	15,101 (52.3)
Other	720 (2.5)	702 (2.4)
Unknown	2,696 (9.4)	2,816 (9.8)
n (%) With 1996 events		
Macro- and microvascular	1,997 (6.9)	1,980 (6.9)
Infectious	810 (2.8)	770 (2.7)
Metabolic	187 (0.6)	129 (0.4)

Descriptions of the best models

For each complication, predictors are shown in the order of entry into stepwise models (Table 2), with ORs for each level of the predictor, and numerical scores assigned to levels that differed significantly from the referent group. Prior hospitalizations (during 1994–1995) for similar events were the strongest predictors of both infectious and metabolic complications and the second strongest predictor of macro- and microvascular complications. Related outpatient diagnoses were the strongest predictor of macro- and microvascular events and were also strongly predictive for infectious complications. There were no outpatient diagnoses for metabolic complications. Increasing age was the third predictor to enter macro- and microvascular and infectious complication models; age was inversely related to metabolic complications.

Several clinical predictors were common to two or all three complications sets. Use of insulin alone (i.e., without records of oral hypoglycemic agents) was associated with increased risk for all three complications sets. Hyperglycemia (average HbA_{1c} level >10.0%), not having HbA_{1c} measured during the baseline period, and elevation of total or LDL cholesterol were all associated with both macro- and microvascular and metabolic complications. Elevated serum creatinine level predicted both macro- and microvascular and infectious disease complications. Outpatient macro- and microvascular disease diagnoses were also a strong predictor of infectious disease events. Use of two or more different antihypertensive medications during the baseline period was a strong predictor of macro- and microvascular events. Interestingly, not having had an albuminuria/microalbuminuria screening, as well as the presence of microalbuminuria or albuminuria, predicted macro- and microvascular events.

Comparisons of the best model with simpler approaches

For macro- and microvascular complications, selection of subjects on the basis of a previous event or related outpatient diagnosis (i.e., the first two variables to enter the model) was as efficient as using the best model, targeting essentially the same proportion of subjects and identifying exactly the same proportion (72%) of 1996 events (Table 3). For infectious and metabolic complications, a prior-events strategy identified far fewer subjects who would have had complications during 1996 than targeting the top 30% of subjects based on model-derived risk scores. However, prior-events strategies, as assessed by positive predictive values, were more efficient because far fewer than 30% of the population was targeted. Not surprisingly, the simple three-variable models, which included previous events and related diagnoses, also did nearly as well as full models, especially for macro- and microvascular complications. Comparisons

Table 2—*Predictors and ORs from the best models predicting 1996 macro- and microvascular, infectious, and metabolic events, derivation data set*

M/M events (*n* = 1,997)			ID events (*n* = 810)			MET events (*n* = 187)		
Predictor	OR	Numerical score	Predictor	OR	Numerical score	Predictor	OR	Numerical score
Outpatient M/M diagnoses (1994–1995)			Inpatient events (1994–1995)			Inpatient MET events (1994–1995)		
			No	1.00				
No	1.00		Yes	2.64*	3	No	1.00	
Yes	2.70*	3	Outpatient M/M diagnoses (1994–1995)			Yes	6.90*	3
Inpatient M/M events (1994–1995)						Diabetes treatment		
			No	1.00		Diet only (reference)	1.00	
No	1.00		Yes	1.45*	1	Oral agents only	0.40†	−3
Yes	1.70*	2	Age (one decade)	1.34*	‡	Insulin and oral agents	0.52	
Age (one decade)	1.24*	‡	Number of visits to specialists			Insulin only	1.60†	2
Antihypertensives			None (reference)	1.00		Mean HbA$_{1c}$ (1994–1995)		
None (reference)	1.00		1–3	1.09		<7.0% (reference)	1.00	
One	1.21†	1	4–6	1.31		7–8%	1.29	
Two or more	1.58*	2	≥7	1.67*	2	9–10%	1.93	
Serum creatinine			Serum creatinine			≥10%	5.07*	3
<1.0 mg/dl	1.00		<1.0 (reference)	1.00		Missing	2.29	
1.0–1.3	1.16		1.0–1.3	1.13		Age (one decade)	0.81*	‡
1.3–1.5	1.24		1.3–1.5	1.74§	2	Emergency department visit (1994–1995)		
1.5–2.0	1.46§	1	1.5–2.0	1.85*	2			
>2.0	1.82*	2	>2.0	2.49*	3	No	1.00	
Missing	0.86		Missing	0.78		Yes	2.04§	3
Diabetes treatment			Outpatient ID diagnoses (1994–1995)			Obesity status		
Diet only (reference)	1.00					Lean (reference)	1.00	
Oral agents only	1.09		No	1.00		Obese	0.39*	−3
Insulin and oral agents	1.06		Yes	1.81*	2	Morbidly obese	0.21*	−3
			Nonmaternity hospitalizations (1993–1995)			BMI missing	1.04	
Insulin only	1.42*	1				Race		
Mean HbA$_{1c}$ (1994–1995)			No	1.00		White (reference)	1.00	
			Yes	1.37§	1	Black	1.00	
<7.0% (reference)	1.00		Treatment			Hispanic/Latino	0.67	
7–8%	1.11		Diet only (reference)	1.00		Asian	0.22†	−3
8–10%	1.33§	1	Oral agents only	1.02		Native American/Other	1.00	
≥10%	1.70*	2	Insulin and oral agents	1.41†	1	Missing	0.90	
Missing	1.22		Insulin only	1.47§	1	Inpatient M/M events (1994–1995)		
Albuminuria								
Absent	1.00					No	1.00	
Present	1.25§	1				Yes	1.76†	2
Missing	1.24§	1				Sex		
Primary care visits						Female	1.00	
None (reference)	1.00					Male	0.61†	−2
1–3	1.13					Smoking status		
4–6	1.17					Nonsmoker (reference)	1.00	
>7	1.43†	1				Ex-smoker	1.02	
Outpatient diagnosis of obesity						Current smoker	1.58‖	2
						Missing	0.72	
No	1.00					Use of antilipemic medications		
Yes	0.73§	–						
Outpatient ID diagnoses						No	1.00	
No	1.00					Yes	0.48§	−3
Yes	1.25§	–1				Mean LDL cholesterol		
Mean total cholesterol						<160 mg/dl (reference)	1.00	
<240 mg/dl (reference)	1.00					≥160 mg/dl	2.02†	3
≥240 mg/dl	1.27§	1				Missing	1.04	
Missing	1.07							

Continued on following page

Table 2—*Continued*

M/M events (n = 1,997)			ID events (n = 810)			MET events (n = 187)		
Predictor	OR	Numerical score	Predictor	OR	Numerical score	Predictor	OR	Numerical score
Self-report of neuropathy								
No	1.00							
Yes	1.26§	1						
Missing	1.06							
Education								
<9 years	1.11							
9–11 years	0.89							
High-school graduate	1.00							
Some college	0.99							
College graduate	0.97							
Graduate school	0.76†	−1						
Missing	1.08							
Type of diabetes								
Type 2 (reference)	1.00							
Type 1	0.64†	−2						
Uncertain	0.71							
Missing	1.02							
Sex								
Female	1.00							
Male	1.13							

*$P < 0.0001$; †P < 0.01; §$P < 0.001$; ‖$P < 0.05$; ‡age was treated as a continuous variable and scored as follows: macro- and microvascular model 19–29 years = 0, 30–39 years = 1, 40–59 years = 2, ≥60 years = 3; infectious disease model 19–29 years = 0, 30–39 years = 1, 40–49 years = 2, ≥50 years = 3; metabolic model 19–29 years = 0; 30–39 years = −1, 40–59 years = −2, ≥60 years = −3; ID, infectious disease; MET, metabolic; M/M, macro- and microvascular.

of ROC curves between full and three-variable models revealed significant differences ($0.01 < P < 0.06$) for each, but differences in the AUCs were quite small (≤4%) for each, suggesting that measurement and inclusion of the remaining variables in the best models adds little to predictive ability.

Simpler numerical scores performed nearly as well as risk scores calculated directly from coefficients of the best models for each complications set. ROC curve comparisons did not reveal any significant differences in AUCs between these two scores (for each complication, $P > 0.05$). All ROC curve comparisons are available upon request (J.V.S.).

An approach based on selecting subjects solely on the basis of elevated HbA_{1c} levels was far less efficient for each complication, whether evaluated at the upper 30% cut point or across the entire range using ROC curve comparisons.

Utility of risk scores in subjects without prior events

Having demonstrated the importance of targeting subjects with previous events or related diagnoses, we compared the remaining approaches in the reduced population of subjects without such markers (Table 4). The first three variables to enter the best model were an elevated serum creatinine level (three levels differed significantly from the reference group of <1.0 mg/dl) followed by use of antihypertensive agents (either one or more than one) and use of insulin as the only therapy. Other significant predictors included a prior emergency department visit, having more than seven primary care visits in the 2-year span, being a current or former smoker, having more than seven outpatient visits to specialists, an average HbA_{1c} level >10.0%, albuminuria or microalbuminuria, and not having microalbuminuria measured during the 2-year interval.

Cumulative sensitivity for 1996 events across the full range of each risk score is shown in Fig. 1. Model sensitivities were not as high in this patient subgroup as in the full sample because of the absence of the two strongest predictors (prior events and related diagnoses). Nevertheless, all three model-based approaches improved substantially over targeting based on HbA_{1c} alone. The numerical score is shown as a black line because its seven observed scores do not fall at decile cut points. There was essentially no difference in performance between the best model and the numerical score as judged by comparison of ROC curves ($P = 0.24$). The AUC for the full model was slightly greater than the AUC for the three-variable model (64 vs. 61%, $P = 0.03$). Identifying patients simply on the basis of their previous HbA_{1c} levels did little better than chance in identifying those at high short-term risk.

CONCLUSIONS — It is frequently observed that very small proportions of a population consume a large fraction of total health costs. In this diabetic population, 20% of the members accounted for 79% of the excess costs of care in 1995 (J.V.S., unpublished data), much of which was a result of hospitalization for complications (2). We aimed to develop a tool that could help to identify those members of a population at greatest risk for complications.

A relatively short-term (1 year) follow-up period was used in these analyses, because decision makers who fund expensive disease management programs

Table 3—*Sensitivity and predictive values of various targeting strategies, validation data set*

Complication	Proportion of population targeted	Sensitivity for 1996 events	Positive predictive value
Macro- and microvascular			
Top 30%* from the "best" model	30	72	16.4
Prior events or diagnoses	31	72	15.8
Top 30%* from three-variable model	30	71	16.1
Top 30%* from numerical score	33	74	15.3
Top 30% of 1994–1995 HbA_{1c} levels	30	31	7.1
Infectious disease			
Top 30%* from the "best" model	30	72	6.4
Prior events or diagnoses	15	44	7.5
Top 30%* from three-variable model	30	68	6.1
Top 30%* from numerical score	30	67	6.0
Top 30% of 1994–1995 HbA_{1c} levels	30	38	3.4
Metabolic			
Top 30%* from the "best" model	30	83	1.2
Prior events or diagnoses	1.5	33	8.5
Top 30%* from three-variable model	30	75	1.0
Top 30%* from numerical score	29	82	1.3
Top 30% of 1994–1995 HbA_{1c} levels	30	59	0.9

Data are %. *Patients with the highest 30% of predicted risk scores in the validation data set. For the numerical risk score, the proportion selected may deviate slightly from 30% because its seven observed values did not allow categorization by deciles.

are highly sensitive to short-term financial considerations (11). Pronk et al. (12) have shown that elevated risk factor levels translate to increased costs for diabetic patients in the short term, and two recent trials of intensive interventions for diabetes (5,13) have shown that hospitalization rates and costs of care can be reduced within 12 months. However, the major predictors in our models (hypertension, hyperglycemia, elevated serum creatinine, use of insulin only, albuminuria, and dyslipidemia) are highly consistent with previous epidemiological (14–16) and intervention studies (17–23) that used a long-term perspective.

Several aspects of the findings should be highlighted. The importance of secondary prevention is demonstrated by the very strong predictive power of prior complications and related outpatient diagnoses. Patients with one or both of these markers accounted for well over half of the complications in 1996 and should clearly be among the first targeted by population disease–management programs. More complex prediction scores, such as those developed here, would be most helpful for targeting primary prevention in the remaining 60–70% of the diabetic population. HbA_{1c} levels predicted increased risk for each set of complications, but model-based targeting improved substantially on selection that was based on elevated HbA_{1c} levels. The simple numerical score, which proved to be as accurate as the score calculated directly from the best-model coefficients, would be the most convenient approach to applying our findings in other populations. In our data, a score ≥7 identified 46% of subjects without prior complications and 66% of their complications in 1996.

Our analyses also indicate that sufficient information for predicting complications is captured in a very small number of commonly available variables. Even among patients with no prior events or related diagnoses, models containing just three variables were nearly as efficient as much more complex models in predicting

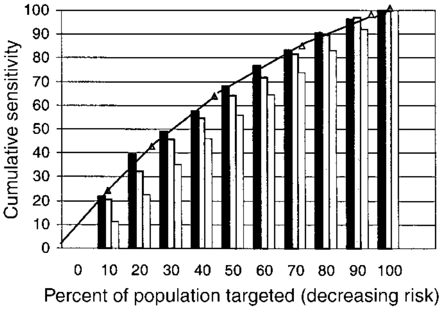

Figure 1—*Sensitivity for any 1996 event (macro- and microvascular, infectious, or metabolic) in the validation data set.* ■, *Decile of risk scores from full model;* ▨, *reduced three-variable model;* □, *average 1994–1995 HbA_{1c} levels for subjects with no prior events or related outpatient diagnoses and serum creatinine <2.0;* —▲—, *numerical risk score.*

Table 4—*Significant predictors of any 1996 event, numerical score, and prevalence of predictor for the derivation sample (restricted to subjects without prior events or related outpatient diagnoses and serum creatinine <2.0 mg/dl)*

Predictor	OR*	Numerical score	Prevalence of predictor (%)
Elevated serum creatinine (mg/dl)			
1.0–1.3	1.40†	1	17.2
1.3–1.5	1.51‡	1	3.1
1.5–2.0	2.66§	3	3.6
Use of one antihypertensive medication	1.45§	1	30.3
Use of ≥2 antihypertensive medications	1.71§	2	31.2
Use of insulin only	1.51§	2	25.6
Emergency visit (1994–1995)	1.36†	1	44.1
≥7 Primary care visits: (1994–1995)	1.36†	1	20.8
Current cigarette smoker	1.49†	1	8.7
≥7 Ambulatory specialist visits	1.30†	1	24.7
Elevated HbA_{1c} (>10%)	1.30‖	1	18.3
Albuminuria (micro- or macro-)			
Present	1.32‖	1	18.5
Missing	1.42†	1	34.7

*Odds ratios use the same referent groups as in Table 2; †$P < 0.01$; ‡$P < 0.05$; ‖$P < 0.0001$.

short-term risk. Nearly all key variables come from data sources (hospital discharge files, outpatient visit claims, laboratory results, and pharmacy records) that are commonly available in health care systems.

Several limitations of these analyses should be kept in mind. First, these risk scores were developed for use by programs that aim to support rather than replace clinical judgment. Although the models confirm the importance of several known clinical risk factors, the model scores derived from automated data are neither sufficiently accurate nor sufficiently complete to supplant decision-making by physicians who treat individual patients in clinical settings. Other information available to the clinician, such as comorbidities or known noncompliance, could easily overrule score-based decisions. The high levels of missing predictor information in our clinically derived data would be considered a serious limitation in epidemiological analyses. However, our aim was to produce a disease-management tool applicable to all members of a population rather than a biological or epidemiological model of complications. By including "missing" as a value for several predictors, we also learned that "missingness" itself can sometimes signal increased risk. We repeated the best models from Table 2, excluding patients with any missing values. Although sample size dropped by as much as 75%, results were essentially identical for macro- and microvascular and infectious models. The metabolic-events model was not interpretable, because the number of end points dropped to 27. Although age was a strong predictor and sex was a weak but significant predictor in at least one of the best models (Table 2), we included neither variable in the final model, because it would make little sense to target only the oldest patients or only one sex for disease management activities.

In conclusion, automated data available in many HMOs could be used to more efficiently identify diabetic patients at high risk for complications. As databases derived directly from electronic medical records replace current systems, the precision and completeness of many predictors will improve, which will further add to the accuracy of predictive models.

Acknowledgments— This research was supported in part by a grant from Pfizer Pharmaceuticals. The authors wish to thank Lyn Wender for her editorial assistance.

APPENDIX

Table A1—*Hospital discharge diagnoses (and International Classification of Diseases-9 Codes) in Each Complications Set*

Diagnosis	Individuals (*n*) with ≥1 events in 1996	International Classification of Diseases-9 Codes
Macro- and microvascular events		
Myocardial infarction	702	410
Other ischemic heart disease	1324	411–414
Coronary artery bypass surgery	162	36.10, 36.11, 36.12, 36.13, 36.14, 36.15, 36.16, 36.19
Percutaneous transluminal angioplasty	58	36.01, 36.02, 36.05
Congestive heart failure	910	428, 402.01, 402.11, 402.91
Cerebrovascular accident	783	431, 433, 434, 436
Chronic renal failure	113	250.4, 585, 586
Lower extremity amputation	381	84.10–84.17
Peripheral vascular disease	501	250.7, 440, 441, 442, 443.9
Gangrene and lower-limb ulcer	41	040.0, 440.23, 440.24, 707.1, 892.1, 785.4
Diabetic eye disease	38	250.5, 362.0, 379.23
Metabolic complications		
Diabetic ketoacidosis	251	250.1
Hyperosmolar coma	28	250.2
Other diabetic coma	13	250.3
Hyperglycemia	31	250.0
Hypoglycemia	240	251
Infectious complications		
Pneumonia	579	480.0–487.8
Septicemia	427	038.0–038.9
Acute pyelonephritis	49	590.1
Chronic pyelonephritis	2	590.0
Renal and perinephric abscess	1	590.2
Other pyelonephritis	1	590.9
Bacteremia	5	790.7
Endocarditis	13	421.0
Osteomyelitis	41	730.0–730.2
Cellulitis and abscess	222	682.0–682.9
Necrotizing fasciitis	7	728.86
Diabetic gangrene	305	250.7
Gangrene (any site)	3	785.4
Gas gangrene	2	040.0
Emphysematous cholecystitis	9	575.0
Fourmier's gangrene	1	608.83
Mucormycosis	1	117.7

References

1. Rubin RJ, Altman WM, Mendelson DN: Health care expenditures for people with diabetes mellitus. *J Clin Endocrinol Metab* 78:809A–812A, 1992
2. Selby JV, Ray GT, Zhang D, Colby CJ: Excess costs of medical care for patients with diabetes mellitus in a managed care population. *Diabetes Care* 20:1396–1402,

Table A2—***Predictor variables examined in stepwise regression analyses in derivation data set***

Variable	Unit of analysis
Membership database	
Patient age	1 year
Patient sex	Male/Female
Inpatient events (1994–1995)*	
Macro- and microvascular complications	Yes/No
Infectious complications	Yes/No
Metabolic complications	Yes/No
Related outpatient diagnoses (1994–1995)	
Macro- and microvascular diagnosis†	Any diagnosis/No diagnosis
Ischemic heart disease	
Renal failure	
Nephropathy	
Cerebrovascular disease	
Peripheral vascular disease	
Gangrene/ulcer of lower extremity	
Proliferative retinopathy	
Photocoagulation treatment	
Infectious diagnosis‡	Any diagnosis/No diagnosis
Abscess/Cellulitis	
Diabetic gangrene	
Laboratory measures‡	
Average HbA_{1c} level	<7%, 7–8%, 8–10%, ≥10%, or missing
Serum creatinine	<1.0 mg/dl, 1.0–1.3, 1.3–1.5, 1.5–2.0, ≥2.0, or missing
Albuminuria or microalbuminuria	Present, absent, or missing
Total cholesterol	≤240 mg/dl, >240, or missing
LDL cholesterol	≤160 mg/dl, >160, or missing
HDL cholesterol	≥35 mg/dl (males) or ≥45 mg/dl (females), <35 or <45 mg/dl, or missing
Triglycerides	≤200 mg/dl, >200 mg/dl, or missing
Ratio of total cholesterol to HDL cholesterol	≤5.6 (females) or ≤6.4 (males), >5.6 or 6.4, or missing
Pharmacy indicators (1994–1995)	
Diabetes treatment	Oral hypoglycemics, insulin, insulin and oral hypoglycemics, no medication
Use of antihypertensive medications	None, one, or more than one
Use of antilipemic medications	Yes/No
Other outpatient diagnoses (1994–1995)	
Peripheral neuropathy	Yes/No
Obesity	Yes/No
Hypertension	Yes/No
Cigarette smoking	Yes/No
Measures of health care use in (1994–1995)	
Number of other hospitalizations	Number
Number of primary care visits	0, 1–3, 4–6, ≥7
Number of urgent care visits	0, 1–3, 4–6, ≥7
Number of emergency department visits	0, 1–3, 4–6, ≥7
Number of ophthalmology/optometry visits	0, 1–3, 4–6, ≥7
Number of other specialty visits	0, 1–3, 4–6, ≥7
Items from patient questionnaire‡	
Race	White, African-American, Hispanic, Asian/Pacific Islander, Native American, other, or missing
Education	<9 years, 9–11 years, high school graduate, some college, college graduate, graduate school, or missing
Annual income	<$10K, $10–20K, $20–40K, ≥40K, or missing
Type of diabetes	Type 1, type 2, uncertain, or missing
Duration of diabetes	<5, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, ≥35 years, or missing
Self-monitoring of blood glucose	Never, less than daily, at least daily, or missing

Continued on following page

Table A2—*Continued*

Variable	Unit of analysis
BMI	<27.3 kg/m^2 in females or <27.8 in males, 27.3–35, >35, missing
Symptoms of neuropathy	Yes/No
History of hypertension	Yes/No
Cigarette smoking status	Current, former, never, or missing
Alcohol consumption	Never, former, or current drinker of <7, 7–13, 14–20, or ≥ 21 drinks/wk, or missing

*See Table A1 for the diagnoses included in each complication set. †Any of the outpatient diagnoses listed, if noted during 1994–1995, were counted as a related outpatient diagnosis for the specific complication set. There were no related outpatient diagnoses applicable to metabolic complications. ‡Missing values are a result of tests not being performed (for laboratory values) and nonresponse for questionnaire items.

1997

3. Peters AL, Davidson MB: Management of patients with diabetes by nurses with support of subspecialists. *HMO Pract* 9:8–13, 1995
4. Aubert RE, Herman WH, Waters J, Moore W, Sutton D, Peterson BL, Bailey CM, Koplan JP: Nurse case management to improve glycemic control in diabetic patients in a health maintenance organization. *Ann Intern Med* 129:605–612, 1998
5. Sadur CN, Moline N, Costa M, Michalik D, Mendlowitz D, Roller S, Watson R, Swain BE, Selby JV, Javorski WC: Diabetes management in a health maintenance organization: efficacy of care management using cluster visits. *Diabetes Care* 22: 2011–2017, 1999
6. Laupacis A, Sekar N, Stiell IG: Clinical prediction rules: a review and suggested modifications of methodological standards. *JAMA* 277:488–494, 1997
7. Selby JV: Linking automated databases for research in managed care settings. *Ann Intern Med* 127:719–724, 1997
8. Centor RM, Keightley J: *ROC Analyzer* (computer software). Washington DC, American Association of Medical Systems and Informatics, 1988
9. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operation characteristic (ROC) curve. *Radiology* 143:29–36, 1982
10. Hanley JA, McNeil BJ: A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology* 148:839–843, 1983
11. Shulkin DJ: Understanding the economics of succeeding in disease management. *Manag Care Interface* 12:98–104, 1999
12. Pronk NP, Goodman MJ, O'Connor PJ, Martinson BC: Relationship between modifiable health risks and short-term health care charges. *JAMA* 282:2235–2239, 1999
13. Rubin RJ, Dietrich KA, Hawk AD: Clinical and economic impact of implementing a comprehensive diabetes management program in managed care. *J Clin Endo Metab* 83:2635–2642, 1998
14. Hypertension in Diabetes Study (HDS). II. Increased risk of cardiovascular complications in hypertensive type 2 diabetic patients. *J Hypertens* 11:319–325, 1993
15. Lehto S, Ronnemaa T, Haffner SM, Pyorala K, Kallio V, Laakso M: Dyslipidemia and hyperglycemia predict coronary heart disease events in middle-aged patients with NIDDM. *Diabetes* 46:1354–1359, 1997
16. Aronow WS: Usefulness of serum creatinine as a marker for coronary events in elderly patients with either systemic hypertension or diabetes mellitus. *Am J Cardiol* 68:678–679, 1991
17. The Diabetes Control and Complications Trial Research Group: The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 329:977–986, 1993
18. UK Prospective Diabetes Study (UKPDS) Group: Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes: UKPDS 33. *Lancet* 352:837–853, 1998
19. UK Prospective Diabetes Study (UKPDS) Group: Tight blood pressure control and risk of macrovascular complications in type 2 diabetes: UKPDS 38. *BMJ* 317: 703–713, 1998
20. Curb JD, Pressel SL, Cutler JA, Savage PJ, Applegate WB, Black H, Camel G, Davis BR, Frost PH, Gonzalez N, Guthrie G, Oberman A, Rutan GH, Stamler J: Effect of diuretic-based antihypertensive treatment on cardiovascular disease risk in older diabetic patients with isolated systolic hypertension. *JAMA* 276:1886–1892, 1996
21. Hansson L, Zanchetti A, Carruthers SG: Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. *Lancet* 351:1755–1762, 1998
22. Pyorala K, Pedersen TR, Kjekshus J, Faergeman O, Olsson AG, Thorgeirsson G: Cholesterol lowering with simvastatin improves prognosis of diabetic patients with coronary heart disease: a subgroup analysis of the Scandinavian Simvastatin Survival Study (4S) *Diabetes Care* 20: 614–620, 1997
23. Koskinen P, Manttari M, ManninenV, Huttunen JK, Heinonen OP, Frick MH: Coronary heart disease incidence in NIDDM patients in the Helsinki Heart Study. *Diabetes Care* 15:820–825, 1992