Grant Final Report

Grant ID: 1R18HS017099-01

Crossing the Quality Assessment Chasm: Aligning Measured and True Quality of Care

Inclusive dates: 09/30/07 - 09/29/10

Principal Investigator:
Mark G. Weiner, MD

Team members:
Diane Richardson, PhD[*]
Elina Medvedeva[*]
Marie Synnestvedt, PhD[†]
John Holmes, PhD[†]
Judith Long, MD[†]
Stan Schwartz, MD[‡]
Sam Field, PhD[‡]
Barbara Turner, MD[‡]
Niyaar Iqbal, MD[§]
Jennifer Garvin, PhD[§]

[*] year 3
[†] years 1-3
[‡] years 1-2
[§] year 1

Performing Organization:
University of Pennsylvania School of Medicine

Project Officer:
Kevin Chaney

Submitted to:
The Agency for Healthcare Research and Quality (AHRQ)
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Abstract

Purpose: Current assessments of quality of care for diabetes rank providers based on the proportion of a provider's panel that achieves specified thresholds for clinical parameters including HBA1c, Blood Pressure and LDL cholesterol. This approach penalizes providers who care for patients who are objectively more difficult to control despite reasonable efforts. The main goal of the project was to develop a model of expected level of control and incorporate this model into a quality measure that ranks providers based on the degree to which their patients are doing better or worse than expected.

Scope: Patients with diabetes having data in EPIC, the Electronic Health Record used by the University of Pennsylvania Health System (UPHS) and VistA, the EHR used by the Philadelphia VA Medical Center.

Methods: We developed models of diagnoses, lab results, vital signs, demographic and visit for their ability to consistently predict A1c, blood pressure, and LDL. Expected values were compared with actual values and provider rankings were calculated based on the degree to which providers were doing better or worse than expected under the different models.

Results: While traditional risk factors of age, gender, and socioeconomic status had some explanatory power in the prediction of control of A1c, we found that the prior degree of control was the most significant predictor of current degree of control for each of the clinical parameters. To be ranked highly, a provider must start with objectively difficult patients having poor control and improve the HBA1c in his patients beyond expectations.

Key Words: electronic health records, quality measurement

Final Report

Purpose

Current quality measures for diabetes control are based on the improved outcomes associated with the year-to-year improvement in Hemoglobin A1c (HBA1c) observed in the intensive treatment arm over traditional care in two major randomized trials completed in the 1990s-- the Diabetes Control and Complications Trial (DCCT)[1] and the UK Prospective Diabetes Study (UKPDS)[2]. Each year, the cross sectional average of the HBA1c in the intensively treated group was better than the corresponding measure in the traditionally treated group, and on average, the intensively treated group had less end organ damage.

In these tightly controlled research settings it was fitting to use cross sectional measures of population level quality improvement. In research settings, enrollment is tightly controlled, interventions are strictly applied, parameters are consistently collected, and dropouts are closely monitored. However, most patient care settings are not so well monitored or controlled. As a result, the translation of this cross sectional measure of quality into actual practice is not simple. It is certainly feasible to collect the most recent HBA1c on patients with diabetes and to compare the proportion of different providers' panels that have patient values below a set threshold. However, this cross sectional approach to quality assessment in what is, inherently, a longitudinal disease in a heterogeneous population, is flawed. In actual clinical practice, patient populations with diabetes are not static, and improvement in the cross sectional quality of care for a population is not necessarily reflected uniformly or consistently over time among the patients who comprise the population. For reasons other than quality of care, patients start out with different baseline degrees of control, move in and out of care settings, see multiple providers, exhibit variability in response to and compliance with medical regimens, may not tolerate medications well, and may not display the steady improvement in control of diabetes that is seen in clinical trials.

This project had three analytical specific aims:

1. Evaluate data structure and clinical issues relevant to the validity of comparisons among providers made using quality measures for diabetes. This aim was a quantitative exploration of the impact of measurement quirks on the assessment of quality. We looked at differences in provider rank when quality measures required patients to achieve certain thresholds clinical values versus change-based measures. We looked at the influence of alterations on the criteria for inclusion. We also explored the impact of integrating data across disparate health systems when patients were seen in both locations

2. Develop a new quality measure for diabetes that accounts for patient heterogeneity in terms of baseline HBA1c and expected trajectory of improvement in diabetes control based on clinical parameters and other data available through the EHR. Current measures of quality of care look only at most recent or average values in the clinical parameters without regard for the prior trajectory of the clinical parameter and the impact of concurrent clinical and demographic issues on the degree of control in that parameter. Studies have shown that non-fatal

comorbidities, notably depression and anxiety, can make it more difficult to achieve optimal thresholds for diabetes control[3,4] because of time constraints limiting the physicians ability to address all the issues the patient brings to the visit.[5] Other comorbidities may present contraindications to the use of certain oral hypoglycemic agents. Even given enough provider-patient time to explore more aggressive treatment strategies in clinically appropriate patients, some conditions such as obesity may render patients more difficult to treat adequately because of higher degrees of insulin resistance. For these and other reasons, some patients continue to have elevated HBA1c despite appropriately increasing doses of oral agents and insulin. In such cases physician effort is reflected in altering medication regimen and other steps that indicate active management of not just the diabetes, but the blood pressure and/or LDL levels. Through this aim we sought to identify patient-related factors associated with poor performance measures that occur despite reasonable attempts by the clinical team to improve the patient's status.

Although we recognize that some characteristics may turn out to be statistically related to a poorer outcome, we are sensitive to the notion of creating a "correction" for these factors. The problem with a correction factor is twofold. First, it is subject to "gaming." For example, if depression is statistically associated with a reduced odds of achieving targets on diabetes quality measures, it is too easy for providers to label other poorly performing patients as having depression, thereby removing these patients from the analysis, or "correcting" for the existence of depression by allowing a different threshold in the clinical parameter to be considered as achieving the quality standard. Second, even without "gaming," the existence of a special analytical status where a different target threshold is set for individuals with certain clinical conditions or demographic characteristics may imply an unwanted and unwarranted double standard in which that those individuals appear less in need of the same degree of diabetes control as others.

As an alternative approach, we sought to develop a quality measure that would assess the degree of actual control relative to expected control in the 3 main clinical parameters associated with diabetes, namely Hemoglobin A1c, Blood Pressure and LDL Cholesterol. Rather than basing a provider's quality ranking on the proportion of his panel achieving a threshold level in the clinical parameter, the ranking would instead be based on the difference between the actual performance and the expected performance. Providers would be rewarded with higher quality scores for having patients that are doing better than expected.

3. Analyze the DCCT data for year-to-year individual variability in diabetes control, particularly within the usual care arm, to assess the impact of variability in an individual's diabetes control over time on microvascular outcomes. For this aim, we purchased the DCCT data which recorded the longitudinal outcomes of the 1441 participants, including rigorous quarterly assessments of HBA1c and important clinical outcomes including nephropathy. We looked at the relative incidence in nephropathy as a function of both average A1c level and A1c variability over the course of the study.

Scope

Current quality standards for the management of diabetes have a strong basis in the medical literature. Data collected between 1983 and 1993 in the DCCT trial provided convincing

evidence that intensive management of Type I diabetes with an insulin pump or 3 or more daily doses of insulin and frequent glucose testing was superior to conventional therapy with twice a day insulin injections. 1441 patients starting with an average hemoglobin A1c (HBA1c) of 8.7% were followed for a mean of 6.5 years, with the intensively managed group achieving an average improvement in HBA1c, with an average HBA1c value below 7% within 6 months that was sustained over the years of the study. The standard of care group actually had a modest worsening of HBA1c over time. Improved clinical outcomes associated with these differences in control were both clinically and statistically significant with a 76% reduction in the risk of primary retinopathy, 69% reduction in the risk of neuropathy and a 34% reduction in the risk of microalbuminuria. The rate of macrovascular outcomes was reduced 41% in the intensive arm, but with the relatively young age of the enrolled population (average age = 27), the baseline incidence of myocardial infarction was expected to be low and the reduction was not statistically significant.

Five years after the publication of the DCCT trial results, the UKPDS published their findings. Their study contained data on 3867 patients with newly diagnosed type II diabetes, collected between 1977 and 1997. About two thirds of the subjects received intensive treatment with insulin or sulfonylureas with a fasting glucose goal of 6 mmol/ l (108 mg/dl). One third of subjects were in a conventionally treated group starting with dietary measures with pharmacological agents added as needed to achieve a goal of a fasting glucose goal of 15 mmol/L (270 mg/dl). As with the DCCT, the intensively treated group in UKPDS shows an initial drop in HBA1c, from 7% to 6%. However, unlike the DCCT, the improvement was not sustained, with a gradual increase in the average HBA1c to over 8% after 10 years. The conventionally treated group also showed a gradual worsening in HBA1c. However, since the conventionally treated group did not experience an early improvement, each year, the conventionally treated group had an average HBA1c value about 1 mg/dl higher than the intensively treated group. Despite the overall rise in HBA1c in both groups, the relatively lower HBA1c in the intensively treated group was associated with 25% fewer microvascular complications and a reduction of 16% (RR= 0.84 95% CI 0.71-1.00) in myocardial infarction. Both groups gained weight over the course of the study, but the intensively treated group experienced greater weight gain (5kg vs. 2.5 kg).

The importance of blood pressure management was also explored as part of another UKPDS analysis,[6] in which hypertensive patients in the groups randomized to receive intensive and traditional glucose management were also stratified to receive “tight” vs “less tight” blood pressure management. By current standards, neither group’s blood pressure targets (150/85 for tight control vs 180/105 for less tight control) were optimal, but still the tight control group had improvements in all stroke and microvascular outcomes and a trend toward improvement in myocardial infarction, regardless of the corresponding degree of glucose control. Other large randomized trials of hypertension including the Hypertension Optimal Treatment (HOT) Trial,[7] the Systolic Hypertension in the Elderly Program (SHEP),[8] the Systolic Hypertension in Europe (Syst-EUR) trial,[9] the Heart Outcomes Prevention Evaluation (HOPE),[10] Losartan Intervention For Endpoint Reduction (LIFE),[11] and Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack (ALLHAT),[12] have supported even more aggressive management of blood pressure in patients with diabetes, culminating in the current JNC 7 recommendation for target blood pressures of <130/ < 80 in this population.

LDL control has also been recognized as a modifiable cardiac risk factor in patients with diabetes. UKPDS 23[13] showed that subjects with an LDL>3.89 had a hazard ratio of 2.26 (1.76-

3.00) of developing CAD and 2.32 (1.41-3.81) for having a fatal myocardial infarction compared with subjects having LDL < 3.02. Subsequent studies[14,15,16,17] have suggested that the presence of diabetes increases the risk of a cardiac event for a person with established coronary artery disease, leading the National Cholesterol Education Program (NCEP) Adult Treatment Panel (ATP) III Guidelines[18] to recommend an LDL target of < 100 for people with diabetes.

Context and Settings

University of Pennsylvania Health System (UPHS). UPHS consists of 3 urban tertiary care inpatient hospitals, totaling about 75,000 admissions/year and a network of ambulatory practices seeing over 2 million visits/year, of which about half are to primary care practices and the other half to subspecialty sites. The primary care sites consist of 5 internal medicine and 1 family medicine practices staffed by Penn faculty, and about 30 community care associates located in southeastern Pennsylvania and southern New Jersey. The Health System is committed to deploy an electronic health record throughout its ambulatory healthcare network and has been working with Epic Systems Corporation, Madison, WI since 1999. Epic's flagship software product, EPIC Hyperspace, is a mature ambulatory EHR with integrated modules designed for physicians, nurses and front and back office staff. It is used for documentation of traditional patient visits as well as phone encounters. Documentation is divided into both discrete fields for information such as vital signs and social history, but also allows for input of unstructured data entry to capture the full detail of a clinical presentation. The system facilitates communication among all members of the healthcare team. All orders for pharmacy, consultations and ancillary tests are written through the system. Results of ordered tests are processed by other Health System information systems that interface with Epic Hyperspace, providing a broad view of patient care activity through a single interface. Epic Hyperspace incorporates health maintenance modules that provide person-specific reminders about the need for essential preventative healthcare activities and provides alerts about potential drug-drug interactions and allergy conflicts.

Currently Epic Hyperspace has been implemented in about 60 ambulatory sites, including all 6 of the faculty-based primary care practices and two of the community care practices. Additionally, EPIC is installed in 3 endocrinology practices having a focus on patients with diabetes. These 8 primary care ambulatory sites, 5 serving an urban, poor community and 3 suburban sites, and the 3 endocrine practices will be the focus of the analysis. Together, the primary care sites saw 63,447 patients in 2006 and 115,910 patients since January, 2001. Twenty-five providers had over 2000 visits while, 45 providers in these practices had over 1000 visits in 2006. Another 160 practitioners consisting of medical residents and part time practitioners saw between 50 and 1000 patients in 2006. There were 4486 patients seen in the endocrine practices in 2006, with the majority of patients seen by 7 physicians. 684 patients were seen in both primary care and endocrine practices.

Through the course of this project, we will identify patients with diabetes according to different definitions based on number and source of diagnosis, laboratory parameters and use of medications. The analysis will include patients seen with diabetes since 2001. A rough analyses reveals that, among the 115,910 patients seen since 2001, 11,088 had at least 1 diabetes diagnosis and 9664 has at least 2 diagnoses. With patient turnover, not all patients are seen every year, but among the 63,447 patients seen in primary care practices in 2006, 7613 patients with at least 1 diabetes diagnosis, and 6,720 patients had at least 2 diagnoses. Among the 4486

patients in the endocrinology practices, 3577 patients had at least one diagnosis of diabetes and 3359 had at least 2 diagnoses. Of the 684 patients common to both endocrine and primary care practices, 610 had at least 1 diagnosis of diabetes and 593 had at least 2 diagnoses

Philadelphia VA Medical Center. The Philadelphia VA Medical Center (PVAMC) is a tertiary care, general medical and surgical hospital located on the western border of the University of Pennsylvania campus. The Philadelphia VAMC provides health care for some 433,000 veterans. The facility is staffed by more than 1,500 employees and supports more than 150 acute beds and a 240-bed Nursing Home Care Unit.
The PVAMC houses a number of primary care and subspecialty clinics. About 16,000 patients are seen annually in primary care clinics. Among this set of patients, 5614 have at least 1 diagnosis of diabetes and 4701 have 2 or more diagnoses.

All the practices within all the VA Medical Centers are now using the Computerized Patient Record System (CPRS), a graphical user interface overlying a MUMPS database and programming environment known as the Veterans Health Information Systems and Technology Architecture (VistA). With its deployment in all departments of all medical centers throughout the VA Health Care System in both inpatient and outpatient settings, VistA encompasses a true longitudinal view of healthcare for veterans. Analogous to Epic Hyperspace, VistA captures ambulatory notes, vital signs, pharmacy and ancillary test orders. VistA, however is broader in scope since it is active in both the inpatient and ambulatory environments. It is also the system used by all the ancillary testing laboratories as well as the hospital inpatient and outpatient pharmacy, so test results and records of pharmacy pick-up behavior are all contained in the same system as the clinical documentation and order entry.

While comprehensive in its data capture, the data are organized in a manner best suited for display of patient information as needed for clinical, and not research purposes. While VistA and CPRS are well suited to display all of a given patient's information, one patient at a time, they cannot be used easily to facilitate analysis of cohorts over time. To overcome this limitation, we created the VA Longitudinal Online Research (VALOR) database, a comprehensive extract of clinically relevant data from VistA organized in a manner suitable for research. As with PICARD, the VALOR database captures data on all clinical encounters, with patient and provider details, visit information including the date and time, diagnoses assigned, procedures performed, and medication and test orders. VALOR also includes detailed data on patients including vital signs and the results of laboratory and ancillary tests.

Methods

Specific Aim 1. Evaluate Data Structure and Clinical Issues Relevant To the Validity of Comparisons among Providers Made Using Quality Measures for Diabetes

The focus of this aim was to decide which patients should "count" as having diabetes for the quality measure and then consider issues in associating patients with the correct provider. This aim also sought to evaluate the impact of including data from external information systems on the identification of patients with diabetes and data that may impact the quality measure.

Who has diabetes? The fundamental focus of the quality measure on diabetes naturally requires a meaningful, reproducible and fair definition of diabetes for the purpose of quality assessment. Consensus on a definition derived partially on the basis of data analysis, but more so on the purpose for which we were creating the definition. We agreed that the definition of diabetes for epidemiological purposes such as measuring the incidence and prevalence of diabetes could and should be different than the definition used for quality assessment. The fundamental definition was the existence of 2 billing diagnoses for diabetes. We ruled out only one billing diagnosis based on the observation that patients with only one billing diagnosis for diabetes had at least a 0.5 mg/dl lower HBA1c than other patients, suggesting that the billing diagnosis was assigned as a rational for ordering the HBA1c test, often to rule out the disease. We gave consideration to the inclusion of patients on a suggestive medication for diabetes even if they did not have any diabetes diagnoses, but ultimately decided against this addition since it was a relatively rare occurrence and often involved metformin which studies have recommended to prevent diabetes in at risk individuals. Since these individuals would tend to have low A1c, we felt inclusion of this group would falsely lower A1c values. We also considered the inclusion of patients found to have elevated HBA1c who were never assigned a diagnosis of diabetes, but decided not to include these patients because of the potential for spurious high readings, and problems of comparison with panels that had more patients with the 2 diagnosis definition, yet were more diet-controlled, and ostensibly easier to control. Still, we felt reports should be generated regarding these patients so that providers can affirm the presence of diabetes for quality reports going forward.

Which provider should be associated with which patient? The raw data available for this aim consists of the visit information including dates of visit activity and the providers and patients involved. Although the EHR does enable the direct association of a patient and provider, we have found that, in many cases, the provider information was not present, or was considered incorrect by the listed provider. Even if the listed provider is correct, it was not clear if the current active provider for the patient should be the one best associated for quality measurement purposes – especially in cases where the provider was recently changed. Other options explored included the provider seen most recently by a patient, and the provider seen most often by the patient in the past 2 years. The provider seen most recently definitely has an association with the patient, though, especially for residents, the most recent provider was likely to be one who was covering for the primary provider. In the end we decided to assign patients, for quality purposes, to providers they saw most often in the past two years. Although this meant that the patient may be associated with a provider they are no longer seeing, it made logical sense to attribute the quality of care to the provider who had the majority of face-to-face contact with the patient.

What is the impact of external data on the assessment of diabetes and its control? In this sub-aim, we hoped to explore the impact of true interoperability of distinct Electronic Health Records at different institutions that may care for the same cohorts of patients. The Philadelphia VA Medical Center (PVAMC) and the University of Pennsylvania Health System (UPHS) are geographically proximate. Many patients seen primarily at the VA also receive subspecialty care at UPHS. The VA requires that for patients to receive free medications under the VA pharmacy plan, they must have a VA primary care provider. Still at least some patients also have primary care providers at UPHS. At the start of this study, the extent of this overlap was unknown. Using a linkage between a database table of demographics and identifiers of VA patients and a

related table of UPHS patients, we were able to find patients with visit activity in both locations and conduct descriptive analysis of their patterns of care and the extent to which A1c, LDL and BP information from the other institution may impact the assessment of quality.

Specific Aim 2. Develop a New Quality Measure for Diabetes That Accounts for Patient Heterogeneity in Terms of Baseline HBA1c and Expected Trajectory of Improvement in Diabetes Control Based on Clinical Parameters and Other Data Available Through the EHR

Through the course of the study, we conducted a number of descriptive analyses of the set of patients meeting our set definition of diabetes that would be difficult to compile using manual methods, but is readily available from EHRs: demographics, diagnoses, heart rate, body weight, body mass index, medication usage, LDL, BP, and A1c values. Without regard to any provider, we looked at population-level trends in A1c over time. We looked at the stability of the clinical parameters in individual patients over time. We explored the degree of concordance in control of LDL, BP and A1c with a goal of deciding if we could develop a meaningful composite measure of quality that accounted for all 3 parameters, or if quality of care would have to be analyzed and reported separately for these parameters. Finally we used clinically informed, data-driven approaches to develop models that consistently predicted current A1c, LDL, and Blood Pressure. We then applied these models to actual patient panels, and created a ranking of providers based on their meeting or exceeding these expectations. Lastly, we conducted a face validity test of the rankings under different assumptions to identify the approach that gives appropriate credit to providers who are truly doing better than their peers, and identifies those that may need more support in order to succeed.

Patient Selection and Study Years. We selected adult patients, older than 18, who had two or more diagnoses of diabetes based on their having a billing ICD9 code in the 250.* range who were seen in one of 10 Primary Care, Internal Medicine, and Diabetic specialty departments. With a focus on UPHS data, all visits and diagnoses between 01/01/2003 and 12/31/2009 were assessed initially to look for the presence of diabetes such that the date of the second diagnosis was taken as the diagnosis date. The duration of diabetes was defined as the time from the diabetes diagnosis date to the most recent visit. We defined physician panel for a given evaluation year as the set of all patients meeting the aforementioned diabetes definition who had at least one visit in that current year, and one visit in the period from 6 months to 2 years before the most recent current-year visit. In this way a patient with diabetes with two quick visits in succession toward the end of an evaluation year and no other visits would NOT be counted. Furthermore we required that included patients had at least two diabetes diagnosis codes before or during the evaluation year and whose duration of diabetes was at least 180 days. Patients whose diabetes was established after the specified year were excluded.

We had planned to conduct analyses on UPHS and VA patients based on an evaluation year of 2007 data and using all data prior to that year for supporting information. Unfortunately, through the course of the analysis we realized there was an error in the measurement of HBA1c that affected all results between March and October 2006. The clinical impact of this error was small – estimated to be 0.3, but from a modeling perspective, this systematic dip in A1c results was significant since it affected the apparent change of many A1c results. We were not comfortable implementing a correction for A1cs drawn in this time period because our

assessment suggested the impact of the lab error was between 0.6 and as high as 1. Therefore we conducted an new analysis based on a 2009 evaluation year where the lab error from 2006 was sufficiently remote to have little impact on the analysis. Unfortunately, we did not have VA IRB approval nor availability of VA HBA1c data to conduct a parallel, contemporaneous analysis at the VA.

Case Mix Model Development. Fundamental to the goals of this project was the need to develop a case-mix model that accounted for baseline differences in patients that may have an independent impact on the outcomes of interest. We explored traditional demographics such as age, race and gender. We tested the impact of weight and BMI, as well as other vital signs like heart rate. We used zip code information correlated with 2000 Census data to provide race-stratified median household income for the zip code as a proxy for patient income. We categorized ICD9-CM codes using the AHRQ Clinical Classification System. Given the large number of CCS categories that our patient set had, we excluded certain categories on the basis of their endogeneity with the outcomes of interest (53, 98, 99, 183, corresponding to lipid and hypertension diagnoses and 49, 50, and 186 corresponding to abnormal glucose diagnoses). We also removed other categories that were of an overly general nature (256 – Medical Exam, 10 – Immunization and screening for infectious disease) as well as upper respiratory infection diagnoses which were universally common. We did not want to include rare categories since, even if the category was closely associated with the outcome, it would at most affect the expected control in a few individuals. Therefore, we required diagnosis categories of interest to have at least 500 patients, though we did include of the set of patients meeting our set definition of diabetes.

For initial case-mix model development and variable selection we considered using the means of A1c (or, BP, LDL) values over the 2 year period, and the log-transformed means, as potential dependent variables. Because we wanted to develop linear models relating patient characteristics to outcomes of care, we examined the distributions of candidate dependent variables. The log-transformed means had distributions that were most suitable for linear regression. Thus, we used the logarithm of the patient's mean A1c as the dependent variable for initial model development.

As independent variables, we considered patient characteristics along with medication use and clinical indicators, summarized below:

- Patient characteristics (9 variables): age, sex, race, income, pulse (number of measures and value), systolic BP (number of measures), LDL (number of measures), HBA1c (number of measures)

- Diabetes Medication use (One for each class of diabetes medication)

- Clinical categories (up to 255 indicators in the CCS system , excluding 10, 53, 98, 99, 183, 49, 50, 186 256 as described in text above)

Therefore, the models that we used for initial variable screening had the general form:

Log(mean A1c (or BP, LDL)) = patient characteristics + medication use + clinical characteristics

which has the general form:

$$\text{Log}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

Then, once we have the values of the β coefficients, we can see their impact on Log(y), for example, when x_1 increases by 1 unit, the log of y increases by the amount β_1.

Initial Model Development. We checked for stability of standard variable selection procedures with our data by running backward and stepwise selection on two independent splits of the data, and then comparing variables selected and coefficient estimates. Since we sought a stable and reliable model that would perform well for case mix adjustment, we then implemented a resampling-based modeling strategy to allow us to explore the variability in selected models more systematically. To do this, we compared models resulting from stepwise selection by running the selection procedure on each of 50 independent random bootstrap samples drawn from the full 2009 panel. We then tabulated the frequency with which each variable was selected across these samples. This allowed us to assess the impact of small changes in the data on model selection, and, as a result, on prediction accuracy and interpretability of the selected models.[19]

Variable Selection: the Bootstrap Lasso Variable Selection Method. For our final model development, we split the panel into equally sized training, test, and validation datasets. We reserved the validation dataset for out-of-sample performance estimation, and used the training and test sets for model development.

In this phase of variable selection, we incorporated resampling directly into the selection procedure, allowing the frequency of selection of each predictor to influence its contribution to the final model, within a framework that has been shown to result in improved model selection consistency.

We chose the Lasso (Least Absolute Shrinkage and Selection Operator)[20,21] rather than stepwise or backward selection, to use in this final phase. Lasso selection improves stability over conventional methods, primarily through imposition of a penalty term during the model fitting that protects against over-fitting on training data. Lasso selection is available in standard statistical software.

The bootstrap Lasso method is a consensus combination approach, based on bootstrap resampling of the data, that performs Lasso selection on each bootstrap sample, then retains only the subset of variables that are consistently selected. This approach leads to consistent and correct variable selection, improving over the single run Lasso in cases of strongly correlated covariates.[22]

In our implementation, we performed Lasso selection on 3000 bootstrap samples and retained the subset of variables that were selected in 90% or more of the samples. We repeated this process for each of three starting seeds for the bootstrap sample selection. We reviewed the small number selected clinical indicators, assessed their explanatory value, and eliminated weak predictors or those that offered little in the way of interpretation.

We retained two selected sets of variables as our 'final models': one (the intersection model) which includes only variables that met the above selection criteria in all three seeds of our bootstrap lasso selection; and the second (the 'union' model), which includes variables that met the selection criteria for any of the three bootstrap seed runs. We assessed model fit using standard diagnostics and performance statistics.

We sought to further improve our approach to take into account provider-level variation in the number of patients, as well as variation in characteristics and measurements inherent in the different patient panels, in estimation of provider mean scores. To this end, we used a simple model-based approach to determine the provider 'effect', or deviation from the overall mean. We entered the provider effect as either a fixed or a random effect in a linear model for the adjusted score, as described below.

Beginning with the patient-level case-mix values of O-E = Z, as the dependent variable in a linear model with provider as the only independent variable, we estimate the provider mean in two ways:

In the fixed effect (FE) estimate, we enter the provider as a fixed independent variable in an ordinary regression model. This gives the same result as computing the mean of the residuals for each provider.

$$Z = \text{Intercept} + \text{Provider}$$

Alternatively, we can estimate the provider mean using models that treat our set of providers as a sample from a larger population of providers. In this framework, we enter the provider as a random effect in a mixed effects model. (ref needed here) and estimate the provider effect within the Empirical Bayes framework. This approach incorporates heterogeneity in within-provider and between-provider variation explicitly, and introduces shrinkage in the estimation of provider mean scores so that estimates for highly variable or very small panel providers are "shrunken" towards the overall mean. A random offset, or intercept term is included for each provider, and the Empirical Bayes(another ref here) estimate is determined from this random provider effect.

Score Based on Current NCQA Recommendations. We compared the above rankings to a ranking of providers based on the % of each provider's panel having current A1c < 8. This is essentially the approach taken by current quality measures for diabetes.

Comparison of Provider Ranks. After the provider-level aggregate or mean score is determined, ranks based on these values are assigned, with lower rank corresponding to more favorable outcomes. Plots comparing provider ranks, based on either unadjusted scores, or ranks from selected case-mix models, provide visual summary of the similarity of the model-based rankings.

Results

Patient characteristics for the 2007 and 2009 patient panels from UPHS are summarized as follows: The 2009 Panel consisted of 6,352 patients, with 56% female and 42.3%, white; the mean age was 62.3, weight 199.8 and the average BMI was 32.2; the mean A1c was 7.3, with a 95% CI from 7.25 to 7.35, in the 2009 panel. The 2007 panel was comprised of 6284 patients, with 57.5% female and 33.5% white, and had similar values for mean age 62.6, weight 200.9, BMI 32.4. The mean A1c in the 2007 panel was significantly lower, 6.9, with a 95% CI from 6.86 to 6.94.

Table 1. Patient characteristics for the 2007 and 2009 patient panels

	UPHS 2007	UPHS 2009	PVAMC 2007
N	6284	6352	4543
%female	57.5	56	1
%white	38	33.5	15
%black	55	60	28
%unknown/other	7	6.5	57
Age (mean)	62.6	62.3	66.8
BMI (Mean)	32.4	32.2	31.4
A1c (mean)	6.9	7.3	7.5
SBP	130	130	134
DBP(mean)	74	74	74

Overlap of Patients between UPHS and the Philadelphia VA Medical Center

Without regard to any specific year, we conducted a coarse analysis of PVAMC patients with at least 1 diagnosis of diabetes in that facility, and matched identifiers against the registration table with the UPHS. We identified 11086 patients within our VAMC who had at least one diagnosis of DM at the PVAMC who were ever seen within UPHS since 1997. Only 3793 of these patients ever had a DM diagnosis recorded at the UPHS, however, many of the UPHS visits were related to cataract procedures and other clinical settings that may be loosely related to DM, but where management of glucose was not the focus. Therefore, we restricted our additional analyses to the 3399 patients seen within primary care or endocrine practices at UPHS. Of this set of patients, only 1126 had an HBA1c ever measured at UPHS. However, of the 3399 patients know to have DM at the VA, only 2281 patients were known to have DM at the UPHS. Of this set, only 1070 patients had an HBA1c at UPHS. Of these patients, 139 had an average HBA1c $\geq$9 and 279 had an average HBA1c $\geq$8. There were 814 patients with HBA1c testing at both institutions. Of the 1211 patients with no HBA1c at UPHS, 965 of them had an HBA1c at the VA; 158 had an average HBA1c >9 and 289 had an average HBA1c $\geq$8.

From these data, we concluded that there were many examples of instances where the measured quality of care for diabetes would have been altered by awareness of the A1c from the other institution, and that in other cases there was unnecessary redundancy in A1c ordering across the institutions.

Composite Versus Distinct Quality Indicators for A1c/LDL/SBP

In view of the importance of simultaneous management of blood pressure, LDL cholesterol, and HbA1c levels in diabetic patients, we considered combining indicators for these individual measures to derive a composite quality indicator whose value might provide a concise summary of outcomes of care. In order to test whether the separate elements to be combined were measuring the same underlying construct, we computed Cronbach alpha statistics for each proposed composite score, and Spearman correlations, ρ, between pairs of measures.

We found only weak correlations between pairs of control indicators, with HbA1c and LDL exhibiting the strongest correlations at 0.086; LDL and blood pressure had ρ=.070; blood pressure control and A1c were the most weakly correlated, with ρ=0.039. The Cronbach alpha values were far below the usual minimum value of .7.

Table 2.

Composite Score	Cronbach Alpha
A1C + BP + LDL	0.167948
A1C + BP	0.074757
BP + LDL	0.120817
A1C + LDL	0.157698

Unlike Greenfield who found that composite measures of outcome across A1c, BP and LDL provided a more stable and meaningful measure of quality, we concluded that our data do not support the use of a single composite measure for diabetes quality of care. Greenfield's study showed a much higher concordance in the level of control across the 3 parameters. Our low concordance meant that one provider could have many patients well controlled on the basis of A1c, while another provider could have many patients well controlled on the basis of SBP. Given this heterogeneity, we felt that the quality of care related to these parameters should be measured and compared separately. Although we initiated independent analyses of all 3 parameters, given the complexity of the modeling process, we only completed a thorough analysis of the A1c.

Exploration of Change in A1c 2007

The following is a graph of the percent change in A1c values over a baseline value set as the average value over the preceding two years.

Figure 1. Percentage of change in A1c values over a baseline value set as the average value over the preceding two years

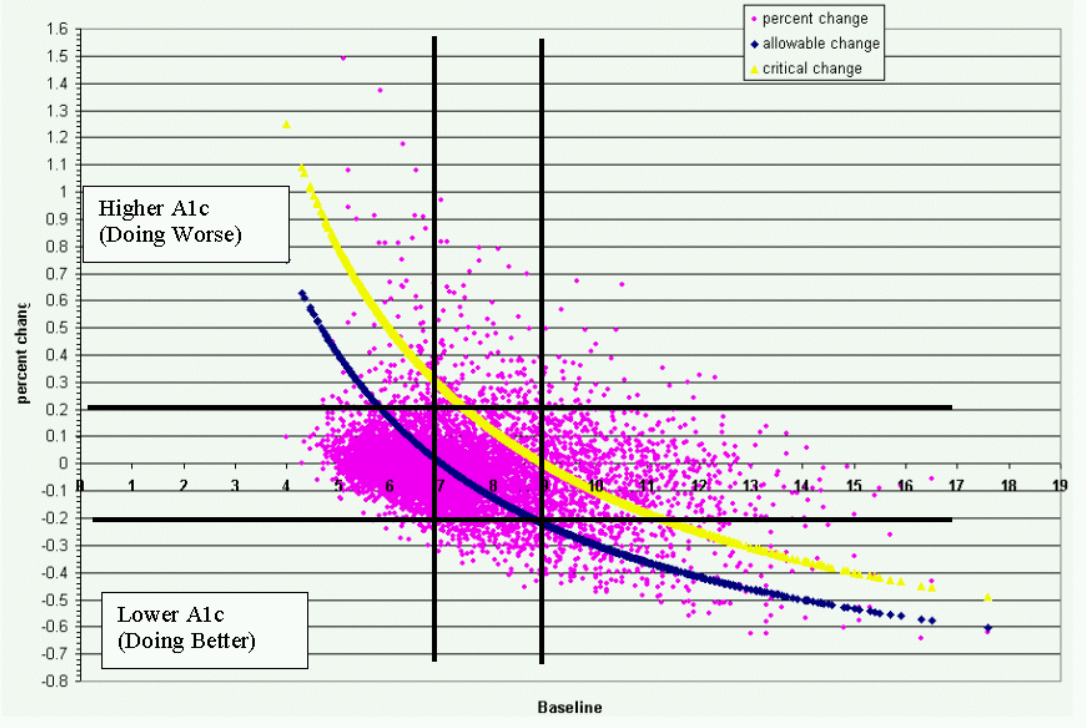


Table 3.

Baseline A1c	<7	7-9	>9
>20% improvement	1%	10%	31%
Within 20% of baseline	96%	84%	65%
>20% worse	3%	6%	4%

The above graph and chart demonstrates that among patients starting with a good A1c defines as a value < 7, the vast majority of patients continue to have a good A1c. While stability of A1c is a dominant feature regardless of baseline A1c, there is a greater degree of change at the higher A1cs. The blue and yellow curves represent the threshold where a patient with a baseline A1c indicated by the x-axis would have a final A1c >7 and >9 respectively.

This finding supported our contention that while having a good HBA1c is desirable clinically, it is relatively common for patients with an already good HBA1c to persist in that state. Patients with initially poor A1c have a greater potential to improve, but a majority remain in their poorly controlled state. This finding suggests that, while better HBA1c control is more desirable, providers who are already caring for patients with good control have an advantage in quality ranking over their peers caring for patients with worse control at baseline. The remainder of the analyses were focused on trying to model the expected degree of HBA1c control as a function of the baseline HBA1c and other clinical and demographic characteristics.

Initial Case Mix Model Selection

Our initial approach to model development for patient-mix adjustment was optimistic, as we sought to select from among a total of over 250 potential variables those that were consistently associated with higher levels of A1c, or, in general, with the less favorable outcome. For initial selection we compared models selected using standard backward and stepwise variable selection on two splits of data; the differences between the selected models confirmed that a more stable selection procedure would be required.

Next, we applied a more sophisticated screening procedure[23] to reduce the starting variable set in a way that did not eliminate variables that showed some predictive strength. We performed stepwise selection beginning with either all or a random subset of clinical indicators, and the non-clinical indicators, on each of 50 bootstrap samples from the 2007 panel. We pooled indicators selected in any of the subset-sample combinations and used this pooled set in a final round of selection using the Lasso.

We repeated this process, varying the sampling starting seed to test whether small perturbations in the procedure would produce different models. We found that there was still variability in the models selected, with a different subset of approximately 20 clinical indicators in each model.

After reviewing the CCS categories of these commonly selected indicators, we arrived at groupings with some plausible face validity, which we have summarized in the appendix as a supplementary table. The groupings included categories which we designated as “benign or symptomatic, but not deadly”, such as: conditions associated with dizziness or vertigo, malaise and fatigue, miscellaneous disorders; and a group of more serious diagnoses, which included categories for liver disease and gastrointestinal hemorrhage. At least in part because of literature supporting an association between depression and related disorders and poor diabetic outcomes,

we considered consolidating indicators in these categories to reduce model complexity while capturing the essence of comorbid conditions that influence patient outcomes.

Interestingly, we noticed that CCS categories for certain more serious diagnoses, which would normally be included in models to adjust for severity, were not selected in any of our screening samples. These included categories representing heart attack, CHF, and various stroke-related categories.

We repeated this procedure for ldl cholesterol and for systolic blood pressure, the two additional quality measures for which we hoped to develop a patient-mix adjustment model. The results were similar, with enough instability in the selection results to preclude a data-driven choice. Effectively, NONE of the CCS diagnostic categories was a stable enough parameter to be used as a case mix adjuster.

However, more traditional patient demographics such as age and income were selected in all models. Additionally the use of Insulin, insulin sensitizing agents (chiefly metformin), and sulfonylureas were predictive such that the patients requiring these medications tended toward higher A1cs. An interesting clinical predictor that appeared in all models was heart rate, such that higher heart rate was associated with a worse A1c.

Given our finding of differential stability of A1c as a function of prior mean A1c, we also tested models that included 3 strata of prior A1c as an independent variable, namely with prior mean A1c values less than 7, between 7 and 9 and greater than 9. Perhaps not surprisingly, this parameter was a strong predictor of current A1c.

Results of Bootstrap Lasso Model Selection

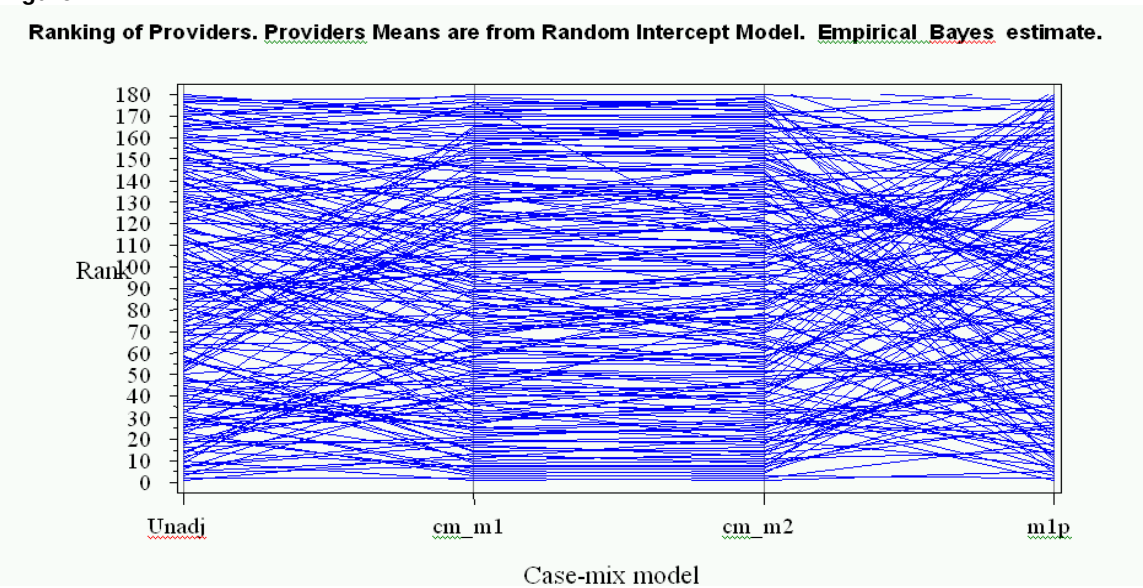
The two models that resulted from the bootstrap Lasso selection were: model M1, which included only variables that were selected consistently in 90% or more of the 3000 bootstrap sample selection runs, across three random starting seeds, and M2, comprised of allows variables that were selected in 90% or more of the 3000 bootstrap selection runs for at least one of the three starting seeds. Six variables were selected for model M1: age, pulse, income, and indicators for the use of insulin, insulin sensitizing agents, and sulfonylureas. The medication group Antidiabetic combinations, and three clinical categories, were selected in model M2, in addition to the six variables from model M1 (198 Other Inflammatory conditions of skin, 212). In view of the strong within-patient correlations of A1c measures, we added a categorical variable for the prior level of A1c control, with levels <7, 7-9, and >9, to model M1 and assessed the fit (using QIC) and stability (using selection probability) of the resulting model, M1p. This model demonstrated improved fit to the data, and the prior A1c level indicator was included with probability close to 1.0. Coefficient estimates for selected models are tabulated in the appendix (or, supplemental tables).

Comparison of Case-Mix Adjusted Provider Ranks

Provider ranks were estimated using each of three selected case mix models and using the unadjusted or intercept-only model. These are plotted on one graph below to permit comparisons. We expect to see fairly large changes between unadjusted and any adjusted ranks. As the figure shows, a number of provider rankings changed moving from the unadjusted model to the M1 model that included age, pulse, income, and indicators for the use of insulin, insulin sensitizing agents, and sulfonylureas. Rankings did not change much between model M1 and

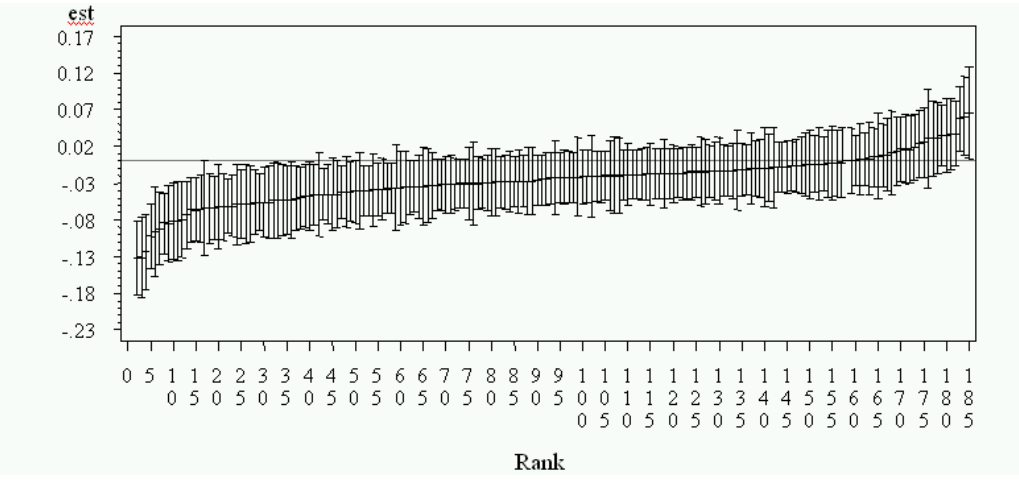
M2, confirming the earlier finding that the inclusion of certain CCS categories in the model would not greatly impact the scores. The most notable feature of the comparison between these sets of ranks is the notable change in ranks associated with model M1p, which includes prior A1c category as well as the other model M1 variables.

Figure 2.



These results suggest that model assumptions can have a strong effect on the measured quality of care where some provider ranking can change drastically depending on the model. This finding is somewhat tempered by the fact that each ranking is based on a point estimate of a score that has some degree of variability associated with it. The following caterpillar plot shows that the confidence intervals around the scores that contribute to the rankings have substantial overlap for all but the most highly and lowest ranked providers:

Figure 3. Provider mean O-E for M1p, fixed provider effect



In spite of the variability in the score that underlies the ranking, analysis of patient panels of providers whose ranking is consistently high, consistently low, or shifts depending on the model provides an interesting face validity assessment that suggests that the M1P model containing age, pulse, income, and indicators for the use of insulin, insulin sensitizing agents, and sulfonylureas as well as the stratification for prior A1c level provides the most credible ranking assessment. For example, one provider would have ranked in the top 10 in terms according to the NCQA approach that looks at the proportion of his panel having an HBA1C <8. In the unadjusted Estimated Bayes model this provider ranked third. However after accounting for clinical and demographic parameters of the provider's panel, this provider's ranking dropped to about 40 out of 185. Finally, using the M1P model, which uses demographic parameters and accounts for prior a HBA1C, this provider's ranking dropped to 117 out of a possible 185. A closer look at the patients belonging to this provider provide a rationale for the lower ranking. Of the 167 patients in this providers panel, there were 139 with a prior a HBA1C drawn. Of these 139 patients, only 16 of them had and a HBA1C > 7. Of the 16 patients within an HBa1c greater than 7, there was a mixture of modest improvement and modest worsening of baseline a HBA1C levels compared to prior values and expected values. The lower ranking does not imply this is a poor quality doctor. Rather, it means he has an already-well-controlled group of patients that he is maintaining under control (which meets with expectations) and has a few poorly controlled patients, some of whom are getting better while others are getting worse. Conversely, another provider ranked in the lower half when assessed in terms of the proportion of the panel with an HBA1c value less than 8. However, this provider was in the top 25% when assessed by the M1P model. Looking at the 24 patients in this panel 42% of the patients has a HBA1c > 8, 10 patients were doing better than expected, while 4 patients were doing worse than expected. The improved ranking assigned to this provider reflects a better-than-expected improvement in HBA1c control among patients starting with poor control. Another provider ranked 46 in terms of the proportion of his panel with an HBA1c < 8, but ranked near last when using the M1P model. In this case 24 of his patients had HBA1c values more than 1 point above the expected value, but only 3 people doing better than expected. His lower score reflects the high proportion of patients doing worse than expected.

Conclusion

Our research supports the notion that the current focus on quality of care for diabetes as measured by proportions of patients in a panel that exceed a quantitative threshold on clinical parameters, in particular HBA1c, is biased, favoring providers with panels that already start with a majority of patients having good control. While some clinical and demographic parameters influence the expectation of control, these factors are dominated by the prior level of diabetes control. We do recognize that effort is certainly expended on the part of both patient and provider to get the HBA1c under better control, but once controlled, the data support an expectation that the HBA1c remain under control. In a panel with many poorly controlled patients, there is a tendency for patients to continue to have poor control, though a certain percentage are expected to show improvement. Under this paradigm where quality of care requires exceeding expectations, the quality of care for diabetes where diabetes control is stable, regardless of the actual level, is always in the middle range. Under this paradigm, providers of patients with poor control have opportunities to demonstrate excellence by having their patients show improvement in their A1c at greater rates or to better thresholds than expectations.

Providers of patients with already good control have less opportunity to demonstrate excellence in quality unless they take on patients with worse control.

Any measure of quality can be criticized for failing to recognize the unmeasurable characteristics of the patients themselves that influence their ability to achieve control. Indeed, we found that many traditional predictors of poor control were not consistently significant in our models. Depression is one notable example. The reason may be related to heterogeneity of the label. Not all patients with depression are equally depressed, some may have responded to treatment, others not. Some may be depressed because of an acute stressful event. For others, it may be more of an ongoing issue that waxes and wanes over time. Discrete data in electronic health records do not express these nuances, and even if the data were contained within the unstructured text of the note, we would not necessarily pick up the patients status between office visits. For these reasons, providers may continue to find fault with these new rankings, just as many find current quality measures problematic, particularly those whose rankings are lowered substantially by the new paradigm.

However, we believe that the baseline A1c, established as an average A1c over 2 years, represents an integration of all patient characteristics, measurable and unmeasurable, that may contribute to the degree of A1c control. Unanticipated problems and new stressors may arise that may impact the pre-existing trend in diabetes control, but our data suggest that the effect of these issues are relatively rare.

Attribution of the level of diabetes control was a difficult problem to address statistically. Were the patients with poor control doing worse because they were seeing objectively poorer providers? It is true that some providers did have a higher concentration of more poorly controlled patients so it is possible that the providers were responsible for the poor control, more than patient behaviors and characteristics. However, even among the providers with poorly controlled patients, in general, their patients who were able to achieve good control did continue to have good control, so that pattern was consistent across all providers.

Our method continues to require refinement. Although we were able to rank providers according to a quantitative score reflecting the degree of diabetes control of their panel with respect to expectations, the standard errors around those scores were large enough where all but the highest and lowest ranked providers were statistically indistinguishable. Part of this is related to the relatively small panel sizes for many of the providers. Future work will focus on expanding the analyses to additional clinical settings to validate the findings of this study.

References

1. DCCT - 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. *NEJM* 1993. 329: 977-86.

2. UK Prospective Diabetes Study 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.

3. Anderson RJ. Grigsby AB. Freedland KE. de Groot M. McGill JB. Clouse RE. Lustman PJ. Anxiety and poor glycemic control: a meta-analytic review of the literature *International Journal of Psychiatry in Medicine.* 32(3):235-47, 2002.

4. Lustman PJ. Clouse RE. Depression in diabetic patients: the relationship between mood and glycemic control. *Journal of Diabetes & its Complications*. 19(2):113-22, 2005.

5. Grant RW. Pirraglia PA. Meigs JB. Singer DE. Trends

in complexity of diabetes care in the United States from 1991 to 2000. *Archives of Internal Medicine*. 164(10):1134-9, 2004.

6. UK Prospective Diabetes Study Group: Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: *UKPDS* 38. BMJ 317:703–713, 1998.

7. Hansson L, Zanchetti A, Camithers SG, Dahlof B, Elmfeldt D, Julius S, et al. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. HOT Study Group. *Lancet*. 1998;351:1755-62.

8. SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA.* 1991;265(24):3255-3264.

9. Staessen JA, Fagard R, Thijs L, Celis H, Arabidze GG, Birkenhager WH, et al. Randomised doubleblind comparison of placebo and active treatment for older patients with isolated systolic hypertension. The Systolic Hypertension in Europe (Syst-Eur) Trial Investigators. *Lancet* 1997;350:757-64.

10. The Heart Outcomes Prevention Evaluation Study Investigators. Effects Of An Angiotensin-Converting–Enzyme Inhibitor, Ramipril,On Cardiovascular Events In High-Risk Patients. *New England Journal of Medicine*. 2000;342:145-53

11. Dahlof B, Devereux RB, Kjeldsen SE, Julius S, Beevers G, Faire U, et al. Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint Reduction in Hypertension Sudy (LIFE): A randomised trial against atenolol. *Lancet* 2002;359:995-1003.

12. ALLHAT Collaborative Research Group. Major cardiovascular events in hypertensive patients randomized to doxazosin vs chlorthalidone: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). *JAMA.* 2000;283(15):1967-1975.

13. Turner RC, Millns R, Neil HAW, Stratton IM, Manley SE, Matthews DR, Holman RR. Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United Kingdom prospective diabetes study (UKPDS: 23). *BMJ* 1998;316;823-828.

14. Haffner SM, Lehto S, Rönnemaa T, Pyörälä K, Laakso M. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. *N Engl J Med* 1998;339:229-34.

15. Haffner SM, D'Agostino RD Jr, Saad MF, O'Leary DH, Savage PJ, Rewers M, Selby J, Bergman RN, Mykkänen L. Carotid artery atherosclerosis in type-2 diabetic and nondiabetic subjects with and without symptomatic coronary artery disease: the Insulin Resistance Atherosclerosis Study. *Am J Cardiol* 2000;85:1395-400.

16. Malmberg K, Yusuf S, Gerstein HC, Brown J, Zhao F, Hunt D, Piegas L, Calvin J, Keltai M, Budaj A, for the OASIS Registry Investigators. Impact of diabetes on long-term prognosis in patients with unstable angina and non-Q-wave myocardial infarction: results of the OASIS (Organization to Assess Strategies for Ischemic Syndromes) Registry. *Circulation* 2000;102:1014-9.

17. Hu FB, Stampfer MJ, Solomon C, Willett WC, Manson JE. Diabetes mellitus and mortality from all causes and coronary heart disease in women: 20 years of follow-up. *Diabetes* 2000;49(suppl 1):A20.

18. National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation*. 2002;106:3143–3421.

19. Harrell Frank E, *Regression Modeling Strategies*: Springer-Verlag, New York 2001

20. Tibshirani, Robert. Regression Shrinkage and Selection via the Lasso, *J. Royal Statistical Society*. Series B, 58(1) 1996, 267-288. See http://www-stat.stanford.edu/~tibs/lasso/lasso.pdf

21. Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. *The Elements of Statistical Learning*, Second Edition, Springer, NY 2009

22. Bach FR, Bolasso: Model Consistent Lasso Estimation through the Bootstrap, Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008

23. Normand, SLT, Glickman, ME, and Gatsonis, CA, Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association*, 92(439), 1997, 803-814