Published in final edited form as:
J Diabetes Complications. 2013 ; 27(6): . doi:10.1016/j.jdiacomp.2013.01.003.

Screening for Impaired Fasting Glucose and Diabetes Using Available Health Plan Data

Laura N. McEwen, PhD[1], **Sara R. Adams, MPH**[2], **Julie A. Schmittdiel, PhD**[2], **Assiamira Ferrara, MD, PhD**[2], **Joseph V. Selby, MD, MPH**[2], and **William H. Herman, MD, MPH**[1,3]

[1]Department of Internal Medicine/Metabolism, Endocrinology and Diabetes University of Michigan, Ann Arbor, MI

[2]Division of Research Kaiser Permanente, Oakland, CA

[3]Department of Epidemiology University of Michigan, Ann Arbor, MI

Abstract

Aims—To develop and validate prediction equations to identify individuals at high-risk for type 2 diabetes using existing health plan data.

Methods—Health plan data from 2005–2009 from 18,527 members of a Midwestern HMO without diabetes, 6% who had fasting plasma glucose (FPG) ≥ 110 mg/dL, and health plan data from 2005–2006 from 368,025 members of a West Coast integrated delivery system without diabetes, 13% who had FPG ≥ 110 mg/dL were analyzed. Within each health plan, we used multiple logistic regression to develop equations to predict FPG ≥ 110 mg/dL for half of the population and validated the equations using the other half. We then externally validated the equations in the other health plan.

Results—Areas under the curve for the most parsimonious equations were 0.665 to 0.729 when validated internally. Positive predictive values were 14% to 32% when validated internally and 14% to 29% when validated externally.

Conclusion—Multivariate logistic regression equations can be applied to existing health plan data to efficiently identify persons at higher risk for dysglycemia who might benefit from definitive diagnostic testing and interventions to prevent or treat diabetes.

Corresponding author: Laura N. McEwen, PhD 1000 Wall Street, Room 6111 Ann Arbor, MI 48105 Telephone: 734-998-6978 Fax: 734-647-2307 lmattei@med.umich.edu.
Laura N. McEwen, lmattei@med.umich.edu
Sara R. Adams, Sara.R.Adams@kp.org
Julie A. Schmittdiel, Julie.A.Schmittdiel@kp.org
Assiamira Ferrara, Assiamira.Ferrara@kp.org
Joseph V. Selby, jselby@pcori.org
William H. Herman, wherman@umich.edu

Competing interests: The authors declare that they have no competing interests.

Authors' contributions: LMN participated in the design of the study, performed the statistical analysis, and drafted the manuscript. SRA performed the statistical analysis and helped draft the manuscript. JAS participated in the design of the study and helped to draft the manuscript. AF participated in the design of the study. JVS and WHH conceived of the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Keywords

screening; impaired fasting glucose; diabetes; administrative data

BACKGROUND

Prediabetes may be defined by a fasting plasma glucose (FPG) ≥110 mg/dL and <126 mg/dL, or a hemoglobin A1c (HbA1c) ≥ 5.7 and < 6.5% (WHO 2006; Ryden, Standl et al. 2007). Prediabetes is associated with an increased risk of type 2 diabetes and cardiovascular disease (Gerstein, Santaguida et al. 2007; Ford, Zhao et al. 2010; ADA 2011). In 2001, the Hoorn Study found that over a six year period, persons with impaired fasting glucose (FPG ≥ 110 mg/dL) were nearly 8 times more likely to develop diabetes than those with normal tolerance at baseline (de Vegt, Dekker et al. 2001). The World Health Organization (WHO) predicts that the worldwide prevalence of diabetes among persons ≥ 20 years of age will increase from 4.0% measured in 1995 to 5.4% by the year 2025 (King, Aubert et al. 1998). To help curb this increase, it is essential to identify persons at high risk for developing diabetes and target them for primary prevention. Interventions are effective in delaying or preventing the development of type 2 diabetes and may reduce the risk of cardiovascular disease (Pan, Li et al. 1997; Tuomilehto, Lindstrom et al. 2001; Chiasson, Josse et al. 2002; Knowler, Barrett-Connor et al. 2002; Gerstein, Yusuf et al. 2006). Unfortunately, diabetes prevention has not been translated into routine clinical practice in part due to the difficulty in identifying individuals at risk.

We hypothesized that within organized systems of care, individuals at risk for type 2 diabetes may be identified without additional laboratory testing. Indeed, high risk individuals may be identified with existing health plan data and targeted laboratory testing may be performed only for those at risk. The goal of our study was to develop and to internally and externally validate equations to screen for prediabetes and previously undiagnosed diabetes using available health plan data. We believe that that the availability of such equations will facilitate the development and widespread implementation of cost-effective interventions to prevent or treat diabetes.

SUBJECTS

To assess the likelihood of impaired fasting glucose (IFG) or previously undiagnosed diabetes, we developed a set of predictive equations using data from a Midwestern independent practice association model health maintenance organization (HMO) that requires primary care physicians to assess and report members' height, weight, blood pressure, FPG, lipids, and smoking status each year. Subjects were at least 18 years of age, not pregnant, had no history of diabetes, and were enrolled in the plan between January 2006 and March 2009. We obtained demographic, claims, and pharmacy data from the health plan, laboratory data from contracted laboratory providers and program enrollment forms, and clinical data from program enrollment forms. Demographic data included age, sex, and race. Claims data included CPT or ICD9 codes for obesity, hypertension, dyslipidemia, gestational diabetes mellitus (GDM), polycystic ovarian syndrome (PCOS), and cardiovascular disease in the 12 months before or after program enrollment (Appendix 1). Pharmacy data included evidence of one or more filled prescriptions for metformin, antihypertensive medications, or lipid-lowering medications in the 12 months before program enrollment (Appendix 1). Laboratory values were obtained for the date closest to enrollment from either the laboratory database or the program enrollment form. Laboratory data included total cholesterol, HDL, triglycerides, and calculated or direct LDL. Clinical

data were obtained from program enrollment forms and included body mass index (BMI), systolic blood pressure, smoking status, and FPG.

Predictive equations were also developed using data from a large West Coast staff model integrated delivery system. Subjects were at least 18 years of age, not pregnant, had no history of diabetes, and had a FPG performed in 2005 or 2006. The plan provided demographic, diagnoses, pharmacy, and laboratory data, as well as clinical data from electronic medical records.

Demographic data included age, sex, and race. Diagnoses included history of obesity, hypertension, dyslipidemia, GDM, PCOS, and cardiovascular disease in the 12 months preceding the date of the FPG test. Pharmacy data included evidence of one or more filled prescriptions for metformin, antihypertensive medications, or lipid-lowering medications in the 12 months before the FPG test date. Laboratory values were obtained from the laboratory database for the date closest to the FPG test and clinical data were obtained from electronic medical records.

For both health plans, we limited the study populations to those ≥ 40 years and < 65 years of age to increase the prevalence of IFG and undiagnosed diabetes and to account for the fact that few subjects ≥65 years of age were enrolled in these commercial health plans. We also excluded from the analyses any person with a filled prescription for a thiazolidinedione in the 12 months before the FPG date.

MATERIALS AND METHODS

Equation development

Data from each health plan were randomly divided into two equal parts, a development set and a validation set. Equations were developed using the development set and internally validated using the validation set.

Using a FPG cutpoint ≥ 110 mg/dL, we classified each subject as having normal glucose tolerance or IFG/diabetes. We then independently developed four equations to incorporate the available data from each health plan. The most simple equation used only demographic and claims or diagnoses data (Equation A). The second used demographic, claims/diagnoses, and pharmacy data (Equation B). The third used demographic, claims/diagnoses, pharmacy, and laboratory data (Equation C). The final equation used demographic, claims/diagnoses, pharmacy, laboratory, and clinical data (Equation D). In the Midwestern HMO, smoking status was available only through clinical data and was included only in Equation D. In the West Coast integrated delivery system, smoking status was available in claims/diagnoses data and was included in Equations B, C, and D.

For each progressively more complex equation, we included only persons with non-missing data for the variables in the equation. For the Midwestern HMO, there were 9,264 individuals included in Equations A and B, 6,439 in Equation C, and 6,116 in Equation D. In the West Coast integrated delivery system, there were 184,197 individuals included in Equations A and B, 149,368 in Equation C, and 138,019 in Equation D. The same populations were used to construct the full and most parsimonious equations (Table 2).

The predictive equations were calculated with the following logistic regression parameters: $P = 1/(1 + e^{-X})$. The final mathematical equation provides an estimate of a subject's likelihood of previously undiagnosed IFG or diabetes expressed as a probability between 0.0 and 1.0. To obtain the most parsimonious equations, we removed variables with p-values > 0.05 in the Midwestern HMO and > 0.01 in the West Coast integrated delivery system.

For each predictive equation, the significance of the variables was assessed by the Wald chi-square test, the estimated odds ratios, and 95% confidence intervals. Using the population with non-missing data for the most complex equation (Equation D), we assessed the fit of each equation by reporting the max-rescaled R^2, and the discrimination of each equation, defined as the ability of the equation to distinguish high-risk subjects as quantified by the area under the curve (C-statistic).

Equation validation

To internally validate the equations, we ran the most parsimonious equations in the health plan in which they were developed using the validation dataset and persons who had non-missing data for each of the variables (Equation D, N=8,682 and N=137,342). To determine the positive predictive value (PPV) of Equations A-D, we used estimates from the most parsimonious equations (Table 2) and applied them to the corresponding validation populations of persons with non-missing data. From these analyses, we estimated the probabilities of IFG or diabetes for each person from each health plan. We then sorted these probabilities into deciles of predicted risk in order from low to high. For each decile of predicted risk, we calculated the percentage of persons with FPG ≥ 110 mg/dL in that decile. This gave us estimates of the PPV for each of the equations.

We then externally validated the equations developed in each health plan using data from the other health plan. For these analyses, we used the same process to calculate PPV, using the entire datasets (development and validation combined) and members who had non-missing data for each variable. The estimates for these equations are presented in Table 3. All analyses were performed using SAS version 9.1.3 or 9.2 (SAS Institute, Cary, NC).

RESULTS

Table 1 describes the characteristics of the populations. Six percent of the subjects in the Midwestern HMO had dysglycemia (N=919 (81%) IFG and N=219 (19%) previously undiagnosed diabetes) and 13% of the subjects in the West Coast integrated delivery system had dysglycemia (N=34,425 (71%) IFG and N=14,243 (29%) previously undiagnosed diabetes). The Midwestern HMO population was younger (mean age 50 years), 78% white, and more likely to be nonsmokers (94%). The West Coast integrated delivery system population was older (mean age 52 years) and more diverse (57% white). Diagnosis and treatment rates differed dramatically between plans but risk factors levels were similar. Substantial proportions of the Midwestern and West Coast populations had claims for hypertension (38% and 27% respectively) and had filled prescriptions for lipid-lowering medications (20% and 18%). In the West Coast integrated delivery system, more people had diagnoses of obesity (27% vs. 11%), however, the average BMI was the same (29 kg/m^2 vs. 29 kg/m^2). In the Midwestern HMO, more people had claims for dyslipidemia (67% vs. 18%), however, the average LDL and HDL levels were similar (LDL: 118 mg/dL vs. 126 mg/dL, HDL: 51 mg/dL vs. 53 mg/dL). In the Midwestern HMO, more persons had claims for cardiovascular disease (21% vs. 5%).

The full equations derived in the development populations are presented in Appendix 2. We have indicated by an asterisk the variables that were included in the most parsimonious equations presented in Table 2 for both the Midwestern HMO and the West Coast integrated delivery system. In the Midwestern HMO, for the equation with the most limited data (Equation A), the most parsimonious equation included older age, male sex, and claims for obesity, hypertension, and dyslipidemia. For Equation B, which also included pharmacy data, the most parsimonious equation did not include dyslipidemia, but instead included lipid-lowering medications and metformin. For Equation C which also included laboratory values, the most parsimonious equation included older age, male sex, claims for obesity and

hypertension, metformin, and higher total cholesterol and lower HDL. For Equation D, which included clinical data, the most parsimonious equation included older age, male sex, race (higher risk for Asians compared to whites), metformin, lower HDL, higher BMI, and higher systolic blood pressure.

When we used the variables in Table 2 and the internal validation population that had non-missing values for all variables in Equation D (N=8,682), the following R^2 and c-statistics were obtained: Equation A: R^2=0.066, C-statistic=0.686, Equation B: R^2=0.089, C-statistic=0.706, Equation C: R^2=0.088, C-statistic=0.706, Equation D: R^2=0.109, C-statistic=0.729 (data not shown in table). The equations explained 6.6% to 10.9%of the variance in the outcome and the areas under the curve were fair, 0.686 to 0.729.

In the West Coast integrated delivery system population, the most parsimonious equation, developed with the most limited data (Equation A) included older age, male sex, and diagnoses for obesity, hypertension, and cardiovascular disease (Table 2). For Equation B, the most parsimonious equation included older age, male sex, race (higher risk for Asian, Other, or missing compared to white), smoking, diagnoses for obesity, hypertension, dyslipidemia, and antihypertensive medications, lipid-lowering medications, and metformin. For Equation C, the most parsimonious equation included older age, male sex, race (higher risk for all groups compared to white), smoking, diagnoses for obesity, hypertension and dyslipidemia, antihypertensive medications, and lipid-lowering medications, lower HDL and higher triglycerides. For Equation D, the most parsimonious equation included older age, male sex, race (higher risk for all groups compared to white), smoking, diagnoses for obesity, hypertension and dyslipidemia, antihypertensive medications, lipid-lowering medications, lower HDL, higher triglycerides, higher BMI, and higher systolic blood pressure.

When using the variables presented in Table 2 and the internal validation population with non-missing values for all the variables in Equation D (N=137,342), the following R^2 and c-statistics were obtained: Equation A: R^2=0.069, C-statistic=0.665, Equation B: R^2=0.083, C-statistic=0.681, Equation C: R^2=0.106, C-statistic=0.703, and Equation D: R^2=0.123, C-statistic=0.718 (data not shown in table).

For each validation population, to determine sensitivity, specificity, and PPV of the most parsimonious equations we sorted the probabilities into deciles of increasing risk and showed the number of persons with IFG or diabetes and the total number of persons in each decile (Table 3). For example, in the Midwestern HMO validation population, for Equation A, the total population was 9,263 and the prevalence of IFG or diabetes was 6% (N=558). In the highest decile of risk, 133 of the 964 persons had IFG or diabetes (PPV = 14%). Sensitivity, the percent of true positives who are identified as such, was 133 / 588 or 24% and specificity, the percent of true negatives who are correctly identified, was 7,874 / 8,705 or 90%. PPV, sensitivity, and specificity were calculated for the remaining equations in a similar manner and Table 4 presents the sensitivity, specificity, and positive predictive values for the highest decide of risk for both the internal and external validation populations. Table 3 can also be used to determine the performance of the equations if one used, for example, the highest three deciles of risk to define a positive test. In the validation population of the Midwestern HMO, if the highest three decides of risk were used to define a positive test, the yield would be (133 + 87 + 58 = 278), and the PPV 10%, the sensitivity 50%, and the specificity 71%.

CONCLUSION

We have demonstrated the feasibility of using existing health plan data to identify individuals at high risk for impaired fasting glucose and previously undiagnosed diabetes

and have documented the performance of equations according to the data available in each health plan. In general, performance improved with the availability of pharmacy, laboratory, and clinical data. Older age and male sex were consistent predictors of IFG or diabetes in all equations and in both health plans. All equations also included one or more indicators of larger body size (either a claim or diagnosis for obesity or a higher measured BMI), blood pressure (either a claim/diagnosis for hypertension, at least one filled prescription for an antihypertensive medication, or systolic blood pressure), and cholesterol (either a claim/diagnosis for dyslipidemia, at least one filled prescription for a lipid-lowering medication, or lipid levels).

Non-white race, smoking, and metformin entered some but not all equations and the effect was plan specific. In the equations developed in the West Coast integrated delivery system, persons of white race had a lower risk of IFG or undiagnosed diabetes than others. In the Midwestern HMO, race entered 2 of 4 equations and persons of Asian race had a higher risk of IFG or undiagnosed diabetes. This may be because the Midwestern population was less diverse resulting in limited power to detect an effect. Smoking was a risk factor in all equations developed for the West Coast integrated delivery system and none of the equations developed for the Midwestern HMO. The West Coast population had a higher proportion of smokers. Since benefits were tied to nonsmoking status in the Midwestern HMO, it is possible that members did not accurately report their smoking status. The true prevalence of smoking in the Midwestern HMO was likely higher than 6%. Finally, metformin was a risk factor in all of the equations developed in the Midwestern HMO but not in Equations C and D developed in the West Coast integrated delivery system. We hypothesize that prescribing patterns for medications for the prevention of diabetes differ between the two health plans.

The equations we developed are similar to those reported by others. Each includes measures of age, obesity, and cardiovascular risk factors (Tabaei and Herman 2002; Tabaei, Engelgau et al. 2005; Kahn, Cheng et al. 2009; Schmid, Vollenweider et al. 2011). However, the previous studies all included patient-reported information available to a clinician such as family history of diabetes, eating, drinking, or physical activity habits. Our equations relied on administrative data available to health plans. Our two most complex equations also included laboratory and clinical data; however, with the increasing use of electronic medical records, these data may become more widely available.

A limitation of the reported equations is their low PPV. PPV is dependent on the prevalence of disease in the population and both of the populations had a low prevalence of IFG or diabetes (6% in the Midwestern HMO and 13% in the West Coast integrated delivery system). The National Health and Nutrition Examination Survey (NHANES) 1999–2002 found the prevalence of IFG or undiagnosed diabetes to be 35.3% in the United States. If 35% of the health plan populations had IFG or undiagnosed diabetes, the PPV for the highest decile of risk for Equation A in the Midwest population would increase from 14% to 56% and in the West Coast population from 26% to 55%. Applying our equations to the available health plan data of less thoroughly screened populations at will have a substantially higher yield.

Previous studies, including randomized controlled clinical trials, have demonstrated that lifestyle or medication interventions can delay or prevent the development of type 2 diabetes in high risk individuals (Pan, Li et al. 1997; Tuomilehto, Lindstrom et al. 2001; Chiasson, Josse et al. 2002; Knowler, Barrett-Connor et al. 2002; Gerstein, Yusuf et al. 2006). These interventions are both effective and cost-effective (Herman, Edelstein et al. 2012). Unfortunately they are not routinely implemented, in part because of the difficulty in identifying individuals at risk. The equations presented here have fairly low sensitivity (20–

25%) but high specificity (~90%). They represent appropriate population screening tests if implemented periodically to identify high risk individuals for definitive diagnostic testing. Screening results can be generated by health plans and reported to members or providers to facilitate diagnostic testing and to trigger interventions. Since these screening equations can be implemented using existing data and without patient or physician involvement, they may provide a low cost tool to improve detection of at-risk populations.

Our goal was to develop flexible equations that could be easily implemented by health systems using existing data in order to identify persons of high risk for IFG or diabetes so that cost-effective lifestyle or medication interventions could be implemented. We have presented eight such equations, each using progressively more complex levels of data that have been validated both internally and externally.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The findings and conclusions in this report are those of the authors and do not represent the views of the funding agencies. Contributions to this study were made by members of the Translating Research into Action for Diabetes (TRIAD) Study Group. The authors acknowledge the participation of our health plan partners. Dr. Selby is currently an employee of the Patient-Centered Outcomes Research Institute (PCORI). The views expressed in this paper are those of the author and do not reflect the views of PCORI.

Grant support: This study was jointly funded by Program Announcement number 04005 from the Centers for Disease Control and Prevention (Division of Diabetes Translation) and the National Institute of Diabetes and Digestive and Kidney Diseases. Support was provided by the Biostatistics Core of the Michigan Diabetes Research and Training Center by Grant Number P60DK020572 and by the Methods and Measurement Core of the Michigan Center for Diabetes Translational Research by Grant Number P30DK09292601 from the National Institute of Diabetes and Digestive and Kidney Diseases. Support was also provided by the NEXT-D study funded by the Centers for Disease Control and Prevention and the National Institute of Diabetes and Digestive and Kidney Diseases. Dr. Schmittdiel was supported by the Health Delivery Systems Center for Diabetes Translational Research (HDS-CDTR) (NIDDK grant 1P30-DK092924).

List of abbreviations

FPG	fasting plasma glucose
HbA1c	hemoglobin A1c
WHO	World Health Organization
IFG	impaired fasting glucose
HMO	health maintenance organization
GDM	gestational diabetes mellitus
PCOS	polycystic ovarian syndrome
HDL	high density lipoprotein
LDL	low density lipoprotein
BMI	body mass index
PPV	positive predictive value

REFERENCES

ADA. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2011; 34(Suppl 1):S62–69. [PubMed: 21193628]

Chiasson JL, Josse RG, et al. Acarbose for prevention of type 2 diabetes mellitus: the STOP-NIDDM randomised trial. Lancet. 2002; 359(9323):2072–2077. [PubMed: 12086760]

de Vegt F, Dekker JM, et al. Relation of impaired fasting and postload glucose with incident type 2 diabetes in a Dutch population: The Hoorn Study. JAMA. 2001; 285(16):2109–2113. [PubMed: 11311100]

Ford ES, Zhao G, et al. Pre-Diabetes and the Risk for Cardiovascular Disease: A Systematic Review of the Evidence. Journal of the American College of Cardiology. 2010; 55(13):1310–1317. [PubMed: 20338491]

Gerstein HC, Santaguida P, et al. Annual incidence and relative risk of diabetes in people with various categories of dysglycemia: a systematic overview and meta-analysis of prospective studies. Diabetes Res Clin Pract. 2007; 78(3):305–312. [PubMed: 17601626]

Gerstein HC, Yusuf S, et al. Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. Lancet. 2006; 368(9541):1096–1105. [PubMed: 16997664]

Herman WH, Edelstein SL, et al. The 10-year cost effectiveness of lifestyle intervention or metformin for diabetes prevention: an intent-to-treat analysis of the DPP/DPPOS. Accepted for publication in Diabetes Care. 2012

Kahn HS, Cheng YJ, et al. Two risk-scoring systems for predicting incident diabetes mellitus in U.S. adults age 45 to 64 years. Ann Intern Med. 2009; 150(11):741–751. [PubMed: 19487709]

King H, Aubert RE, et al. Global burden of diabetes, 1995–2025: prevalence, numerical estimates, and projections. Diabetes Care. 1998; 21(9):1414–1431. [PubMed: 9727886]

Knowler WC, Barrett-Connor E, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002; 346(6):393–403. [PubMed: 11832527]

Pan XR, Li GW, et al. Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. Diabetes Care. 1997; 20(4):537–544. [PubMed: 9096977]

Ryden L, Standl E, et al. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). Eur Heart J. 2007; 28(1):88–136. [PubMed: 17220161]

Schmid R, Vollenweider P, et al. Estimating the risk of developing type 2 diabetes: a comparison of several risk scores: the Cohorte Lausannoise study. Diabetes Care. 2011; 34(8):1863–1868. [PubMed: 21636799]

Tabaei BP, Engelgau MM, et al. A multivariate logistic regression equation to screen for dysglycaemia: development and validation. Diabet Med. 2005; 22(5):599–605. [PubMed: 15842515]

Tabaei BP, Herman WH. A multivariate logistic regression equation to screen for diabetes: development and validation. Diabetes Care. 2002; 25(11):1999–2003. [PubMed: 12401746]

Tuomilehto J, Lindstrom J, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med. 2001; 344(18):1343–1350. [PubMed: 11333990]

WHO. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. 2006

Table 1

Characteristics of the study populations stratified by glucose tolerance: normal glucose tolerance (NGT) vs impaired fasting glucose (IFG) or diabetes (DM) (defined as fasting plasma glucose [FPG] ≥ 110 mg/dL).

	Midwestern HMO			West Coast integrated delivery system		
Characteristic	**NGT N=17,389**	**IFG or DM N= 1,138**	**Total Population N=18,527**	**NGT N=319,357**	**IFG or DM N=48,668**	**Total Population N=368,025**
Demographics						
Age (years) (mean ± std dev)	50 ± 7	53 ± 7	50 ± 7	52 ± 7	54 ± 7	52 ± 7
Sex (N,% female)	9,315 (54%)	431 (38%)	9,746 (53%)	181,459 (57%)	20,912 (43%)	202,371 (55%)
Race (N, %)						
White	13,623 (78%)	911 (80%)	14,534 (78%)	181,430 (57%)	26,752 (55%)	208,182 (57%)
Black	748 (4%)	51 (4%)	799 (4%)	18,514 (6%)	2,980 (6%)	21,494 (6%)
Asian	836 (5%)	53 (5%)	889 (5%)	43,152 (14%)	6,680 (14%)	49,832 (14%)
Other	610 (4%)	30 (3%)	640 (3%)	4,656 (1%)	886 (2%)	5,542 (2%)
Missing	1,572 (9%)	93 (8%)	1,665 (9%)	71,605 (22%)	11,370 (23%)	82,975 (23%)
Diagnoses						
Obesity (N, %)	1,752 (10%)	198 (17%)	1,950 (11%)	78,627 (25%)	19,586 (40%)	98,213 (27%)
Hypertension (N, %)	6,470 (37%)	659 (58%)	7,129 (38%)	78,582 (25%)	20,118 (41%)	98,700 (27%)
Dyslipidemia (N, %)	11,521 (66%)	881 (77%)	12,402 (67%)	54,409 (17%)	11,113 (23%)	65,522 (18%)
Cardiovascular disease (N, %)	3,590 (21%)	306 (27%)	3,896 (21%)	15,858 (5%)	3,991 (8%)	19,849 (5%)
History of GDM (females only) (N, %)	93 (1%)	3 (0%)	96 (1%)	189 (0%)	30 (0%)	219 (0%)
History of PCOS (females only) (N, %)	58 (0%)	1 (0%)	59 (0%)	91 (0%)	14 (0%)	105 (0%)
Medications						
Any antihypertensive medication (N, %)	2,552 (15%)	293 (26%)	2,845 (15%)	103,447 (32%)	25,863 (53%)	129,310 (35%)
Any lipid-lowering medication (N, %)	3,240 (19%)	388 (34%)	3,628 (20%)	52,524 (16%)	12,054 (25%)	64,576 (18%)
Metformin (N, %)	101 (1%)	45 (4%)	146(1%)	61 (0%)	28 (0%)	89 (0%)
Laboratory values						
Total cholesterol (mg/dL) (mean ± std dev)	194 ± 34	195 ± 38	194 ± 35	207 ± 37	209 ± 39	208 ± 37
HDL (mg/dL) (mean ± std dev)	52 ± 15	46 ± 14	51 ± 15	54 ± 15	48 ± 13	53 ± 15
Triglycerides (mg/dL) (mean ± std dev)	122 ± 93	163 ± 109	124 ± 95	143 ± 91	184 ± 113	148 ± 95
LDL (mg/dL) (mean ± std dev)	118 ± 29	119 ± 33	118 ± 30	126 ± 32	126 ± 33	126 ± 32
Clinical data						

Characteristic	Midwestern HMO			West Coast integrated delivery system		
	NGT N=17,389	IFG or DM N= 1,138	Total Population N=18,527	NGT N=319,357	IFG or DM N=48,668	Total Population N=368,025
BMI (kg/m^2) (mean ± std dev)	29 ± 6	33 ± 7	29 ± 6	28 ± 6	32 ± 7	29 ± 6
Systolic blood pressure (mmHg) (mean ± std dev)	121 ± 13	127± 13	121 ± 13	125 ± 15	129± 15	125± 15
Current smoker (N, %)	1,012 (6%)	82 (7%)	1,094 (6%)	38,585 (12%)	7,751 (16%)	46,336 (13%)

GDM: gestational diabetes mellitus

PCOS: polycystic ovarian syndrome

Table 2

Maximum likelihood estimates of logistic regression functions for the most parsimonious version of each equation.

Midwestern HMO					
Variable	**Estimated Regression Coefficient**	**Estimated Standard Error**	**Wald P-value**	**Estimated Odds Ratio**	**95% Confidence Interval for the Odds Ratio**
Equation A (demographics and diagnoses)					
N = 9,264					
C-statistic parsimonious model = 0.667					
R-square parsimonious model = 0.056					
Intercept	−4.7584	0.3359	<0.0001	--	--
Age (years)	0.0364	0.0065	<0.0001	1.04	1.02 – 1.05
Female	−0.7191	0.0909	<0.0001	0.49	0.41 – 0.58
Obesity	0.6633	0.1178	<0.0001	1.94	1.54 – 2.45
Hypertension	0.5399	0.0934	<0.0001	1.72	1.43 – 2.06
Dyslipidemia	0.2222	0.1076	0.0390	1.25	1.01 – 1.54
Equation B (demographics, diagnoses, medications)					
N = 9,264					
C-statistic parsimonious model = 0.680					
R-square parsimonious model = 0.068					
Intercept	−4.5358	0.3413	<0.0001	--	--
Age (years)	0.0338	0.0067	<0.0001	1.03	1.02 – 1.05
Female	−0.7040	0.0922	<0.0001	0.50	0.41 – 0.59
Black (reference=white)	−0.5339	0.2639	0.0430	0.59	0.35 – 0.98
Asian (reference=white)	0.1455	0.2162	0.5010	1.16	0.76 – 1.77
Other race(reference=white)	−0.0970	0.2577	0.7068	0.91	0.55 – 1.50
Race missing (reference=white)	−0.2168	0.1678	0.1963	0.81	0.58 – 1.12
Obesity	0.6301	0.1197	<0.0001	1.89	1.49 – 2.37
Hypertension	0.5289	0.0938	<0.0001	1.70	1.41 – 2.04
Lipid-lowering medication	0.3183	0.0997	0.0014	1.38	1.13 – 1.67
Metformin	1.4865	0.2794	<0.0001	4.42	2.56 – 7.65
Equation C (demographics, diagnoses, medications, and laboratory values)					
N= 6,439					

Midwestern HMO					
Variable	**Estimated Regression Coefficient**	**Estimated Standard Error**	**Wald P-value**	**Estimated Odds Ratio**	**95% Confidence Interval for the Odds Ratio**
C-statistic parsimonious model = 0.692					
R-square parsimonious model = 0.075					
Intercept	−4.6627	0.5041	<0.0001	--	--
Age (years)	0.0470	0.0079	<0.0001	1.05	1.03 – 1.06
Female	−0.4302	0.1189	0.0003	0.65	0.52 – 0.82
Obesity	0.5203	0.1416	0.0002	1.68	1.28 – 2.22
Hypertension	0.4033	0.1120	0.0003	1.50	1.20 – 7.86
Metformin	1.5063	0.3090	<0.0001	4.51	2.46 – 8.26
Total cholesterol (mg/dL)†	0.0031	0.0015	0.0390	1.03	1.00 – 1.06
HDL (mg/dL)†	−0.0246	0.0046	<0.0001	0.78	0.71 – 0.86
Equation D (demographics, diagnoses, medications, laboratory values, and clinical data)					
N=6,116					
C-statistic parsimonious model = 0.733					
R-square parsimonious model = 0.112					
Intercept	−8.9212	0.6903	<0.0001	--	--
Age (years)	0.0481	0.0080	<0.0001	1.05	1.03 – 1.07
Female	−0.4769	0.1249	0.0001	0.62	0.49 – 0.79
Black (reference=white)	−0.6731	0.3554	0.0582	0.51	0.25 – 1.02
Asian (reference=white)	0.6282	0.2441	0.0101	1.87	1.16 – 3.02
Other race (reference=white)	0.1355	0.3023	0.6541	1.15	0.63 – 2.07
Race missing (reference=white)	−0.1754	0.2142	0.4129	0.84	0.55 – 1.28
Metformin	1.2929	0.3267	<0.0001	3.64	1.92 – 6.91
HDL (mg/dL) †	−0.0140	0.0047	0.0030	0.87	0.79 – 0.95
BMI (kg/m^2)	0.0704	0.0079	<0.0001	1.07	1.06 – 1.09
Systolic blood pressure (mmHg) †	0.0195	0.0043	<0.0001	1.22	1.12 – 1.32

West Coast integrated delivery system					
Variable	**Estimated Regression Coefficient**	**Estimated Standard Error**	**Wald P-value**	**Estimated Odds Ratio**	**95% Confidence Interval for the Odds Ratio**
Equation A (demographics and diagnoses)					
N= 184,197					
C-statistic parsimonious model = 0.660					
R-square parsimonious model = 0.066					
Intercept	−3.6199	0.0579	<0.0001	--	--
Age (years)	0.0310	0.0011	<0.0001	1.03	1.03 – 1.03
Female	−0.5766	0.0142	<0.0001	0.56	0.55 – 0.58
Obesity	0.6139	0.0150	<0.0001	1.85	1.79 – 1.90
Hypertension	0.5382	0.0152	<0.0001	1.71	1.66 – 1.76
CVD	0.1047	0.0272	0.0001	1.11	1.05 – 1.17
Equation B (demographics, diagnoses, medications)					
N= 184,197					
C-statistic parsimonious model = 0.667					
R-square parsimonious model = 0.080					
Intercept	−3.8700	0.0621	<0.0001	--	--
Age (years)	0.0301	0.0011	<0.0001	1.03	1.03 – 1.03
Female	−0.5533	0.0145	<0.0001	0.58	0.56 – 0.59
Black (reference=white)	0.0468	0.0300	0.1187	1.05	0.99 – 1.11
Asian (reference=white)	0.3313	0.0217	<0.0001	1.39	1.33 – 1.45
Other race (reference=white)	0.3226	0.0227	<0.0001	1.38	1.32 – 1.44
Race missing (reference=white)	0.2262	0.0198	<0.0001	1.25	1.21 – 1.30
Smoking	0.3003	0.0198	<0.0001	1.35	1.30 – 1.40
Obesity	0.6333	0.0154	<0.0001	1.88	1.83 – 1.94
Hypertension	0.1474	0.0212	<0.0001	1.16	1.11 – 1.21
Dyslipidemia	−0.1327	0.0214	<0.0001	0.88	0.84 – 0.91
Antihypertensive medication	0.5433	0.0207	<0.0001	1.72	1.65 – 1.79
Lipid-lowering medication	0.1813	0.0210	<0.0001	1.20	1.15 – 1.25
Metformin	0.8725	0.3258	0.0074	2.39	1.26 – 4.53
Equation C (demographics, diagnoses, medications, and laboratory values)					
N= 149,368					
C-statistic parsimonious model = 0.701					

West Coast integrated delivery system					
Variable	**Estimated Regression Coefficient**	**Estimated Standard Error**	**Wald P-value**	**Estimated Odds Ratio**	**95% Confidence Interval for the Odds Ratio**
R-square parsimonious model = 0.104					
Intercept	−3.4562	0.0788	<0.0001	--	--
Age (years)	0.0339	0.0013	<0.0001	1.03	1.03 – 1.04
Female	−0.2785	0.0174	<0.0001	0.76	0.73 – 1.78
Black (reference=white)	0.1606	0.0327	<0.0001	1.17	1.10 –1.25
Asian (reference=white)	0.3134	0.0235	<0.0001	1.37	1.31 – 1.43
Other race (reference=white)	0.2440	0.0251	<0.0001	1.28	1.22 – 1.34
Race missing (reference=white)	0.1788	0.0229	<0.0001	1.20	1.14 – 1.25
Smoking	0.2046	0.0227	<0.0001	1.23	1.17 – 1.28
Obesity	0.5183	0.0172	<0.0001	1.68	1.62 – 1.74
Hypertension	0.1141	0.0228	<0.0001	1.12	1.07 – 1.17
Dyslipidemia	−0.1441	0.0220	<0.0001	0.87	0.83 – 0.90
Antihypertensive medication	0.5045	0.0224	<0.0001	1.66	1.58 – 1.73
Lipid-lowering medication	0.1717	0.0215	<0.0001	1.19	1.14 – 1.24
HDL (mg/dL) †	−0.0193	0.0007	<0.0001	0.82	0.81 – 0.84
Triglycerides (mg/dL) †	0.0019	0.0001	<0.0001	1.02	1.02 – 1.02
Equation D (demographics, diagnoses, medications, laboratory values, and clinical data)					
N= 138,019					
C-statistic parsimonious model = 0.720					
R-square parsimonious model = 0.069					
Intercept	−6.4338	0.1168	<0.0001	--	--
Age (years)	0.0377	0.0014	<0.0001	1.04	1.04 – 1.04
Female	−0.3290	0.0183	<0.0001	0.72	0.69 – 0.75
Black (reference=white)	0.0891	0.0339	0.0087	1.09	1.02 – 1.17
Asian (reference=white)	0.5153	0.0250	<0.0001	1.67	1.59 – 1.76
Other race (reference=white)	0.2516	0.0261	<0.0001	1.29	1.22 – 1.35
Race missing (reference=white)	0.2192	0.0245	<0.0001	1.25	1.19 – 1.31
Smoking	0.2561	0.0238	<0.0001	1.29	1.23 – 1.35
Obesity	0.0907	0.0208	<0.0001	1.09	1.05 – 1.14
Hypertension	0.1083	0.0241	<0.0001	1.11	1.06 – 1.17
Dyslipidemia	−0.0702	0.0229	0.0022	0.93	0.89 – 0.98
Antihypertensive medication	0.3751	0.0239	<0.0001	1.46	1.39 − 1.52
Lipid-lowering medication	0.1487	0.0226	<0.0001	1.16	1.11 – 1.21

West Coast integrated delivery system					
Variable	**Estimated Regression Coefficient**	**Estimated Standard Error**	**Wald P-value**	**Estimated Odds Ratio**	**95% Confidence Interval for the Odds Ratio**
HDL (mg/dL) †	−0.0148	0.0008	<0.0001	0.86	0.85 – 0.88
Triglycerides (mg/dL) †	0.0019	0.0001	<0.0001	1.02	1.02 – 1.02
BMI (kg/m2)	0.0635	0.0016	<0.0001	1.07	1.06 – 1.07
Systolic blood pressure (mmHg) †	0.0065	0.0006	<0.0001	1.07	1.05 – 1.08

† for these variables the estimated odds ratio and 95% confidence interval for the odds ratio are per 10 unit increase

Table 3

Positive predictive values of the most parsimonious equations given by number and percent of patients with IFG or diabetes in each decile of predicted risk.

Midwestern HMO Validation population								
Decile	**Equation A**		**Equation B**		**Equation C**		**Equation D**	
	IFG or DM / total	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**
Total	558 / 9,263	6%	558 / 9,263	6%	540 / 9,032	6%	514 / 8,682	6%
1	20 / 877	2%	23 / 838	3%	18 / 903	2%	6 / 868	1%
2	21 / 922	2%	20/ 1011	2%	19 / 903	2%	17 / 868	2%
3	31 / 947	3%	27 / 895	3%	22 / 903	2%	23 / 868	3%
4	35 / 934	4%	35 / 954	4%	27 / 904	3%	28 / 869	3%
5	46 / 948	5%	44 / 933	5%	37 / 903	4%	33 / 868	4%
6	46 / 903	5%	46 / 927	5%	53 / 903	6%	45 / 868	5%
7	81 / 923	9%	69 / 926	7%	51 / 904	6%	50 / 869	6%
8	58 / 943	6%	63 / 925	7%	77 / 903	9%	62 / 868	7%
9	87 / 902	10%	92 / 914	10%	84 / 903	9%	108 / 868	12%
10	133 / 964	14%	139 / 940	15%	152 / 903	17%	142 / 868	16%

West Coast integrated delivery system total population using Midwestern HMO estimates								
Decile	**Equation A**		**Equation B**		**Equation C**		**Equation D**	
	IFG or DM / total	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**
Total	48,668 / 368,025	13%	48,668 / 368,025	13%	42,699 / 321,113	13%	39,471 / 296,693	13%
1	1,515 / 35,302	4%	1,722 / 36,797	5%	950 / 32,110	3%	894 / 29,669	3%
2	2,091 / 35,539	6%	2,222 / 36,595	6%	1,615 / 32,108	5%	1,455 / 29,669	5%
3	2,916 / 36,870	8%	2,891 / 37,008	8%	2,338 / 32,111	7%	2,033 / 29,670	7%
4	3,401 / 36,398	9%	3,571 / 36,750	10%	2,972 / 32,116	9%	2,766 / 29,668	9%
5	4,561 / 39,609	12%	4,305 / 36,834	12%	3,649 / 32,104	11%	3,296 / 29,670	11%
6	4,955 / 36,552	14%	4,793 / 35,634	13%	4,360 / 32,117	14%	3,829 / 29,670	13%
7	5,579 / 36,915	15%	5,619 / 37,973	15%	5,022 / 32,108	16%	4,589 / 29,669	16%
8	6,227 / 37,016	17%	6,219 / 36,510	17%	5,897 / 32,112	18%	5,447 / 29,670	18%

West Coast integrated delivery system total population using Midwestern HMO estimates								
Decile	**Equation A**		**Equation B**		**Equation C**		**Equation D**	
	IFG or DM / total	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**
9	7,794 / 36,581	21%	7,698 / 36,820	21%	6,877 / 32,116	21%	6,561 / 29,669	22%
10	9,629 / 37,243	26%	9,628 / 37,104	26%	9,019 / 32,111	28%	8,611 / 29,669	29%

West Coast integrated delivery system Validation population								
Decile	**Equation A**		**Equation B**		**Equation C**		**Equation D**	
	IFG or DM / total	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**
Total	24,317 / 183,828	13%	24,317 / 183,828	13%	20,166 / 149,244	14%	18,635 / 137,786	14%
1	794 / 18,254	4%	641 / 17,954	4%	406 / 14,923	3%	308 / 13,778	2%
2	1,065 / 18,193	6%	929 / 18,811	5%	686 / 14,925	5%	555 / 13,778	4%
3	1,434 / 18,570	8%	1,353 / 18,146	7%	944 / 14,920	6%	893 / 13,780	6%
4	1,621 / 17,107	9%	1,694 / 18,533	9%	1,272 / 14,929	9%	1,098 / 13,778	8%
5	2,126 / 19,013	11%	2,034 / 18,450	11%	1,664 / 14,922	11%	1,500 / 13,779	11%
6	2,409 / 17,810	14%	2,403 / 18,348	13%	1,980 / 14,927	13%	1,857 / 13,779	13%
7	2,890 / 19,339	15%	2,807 / 18,437	15%	2,347 / 14,925	16%	2,161 / 13,778	16%
8	3,033 / 18,316	17%	3,360 / 18,252	18%	2,974 / 14,924	20%	2,705 / 13,779	20%
9	4,064 / 18,576	22%	4,061 / 18,488	22%	3,451 / 14,925	23%	3,203 / 13,779	23%
10	4,881 / 18,650	26%	5,035 / 18,409	27%	4,442 / 14,924	30%	4,355 / 13,778	32%

Midwestern HMO total population using West coast integrated delivery system estimates								
Decile	**Equation A**		**Equation B**		**Equation C**		**Equation D**	
	IFG or DM / total	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**
Total	1,138 / 18,527	6%	1,138 / 18,527	6%	804 / 13,077	6%	761 / 12,456	6%
1	37 / 1,690	2%	39 / 1,845	2%	27 / 1,305	2%	19 / 1,245	2%
2	49 / 1,997	2%	44 / 1,814	2%	25 / 1,309	2%	19 / 1,246	2%
3	65 / 1,845	4%	50 / 1,876	3%	30 / 1,309	2%	22 / 1,246	2%
4	66 / 1,703	4%	80 / 1,871	4%	41 / 1,307	3%	36 / 1,245	3%
5	85 / 1,883	5%	99 / 1,851	5%	59 / 1,308	5%	49 / 1,245	4%

Midwestern HMO total population using West coast integrated delivery system estimates								
Decile	**Equation A**		**Equation B**		**Equation C**		**Equation D**	
	IFG or DM / total	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**	**IFG or DM / total**	**Percent with IFG/DM**
6	96 / 1,848	5%	99 / 1,859	5%	71 / 1,308	5%	61 / 1,247	5%
7	147 / 1,976	7%	128 / 1,851	7%	101 / 1,308	8%	97 / 1,245	8%
8	145 / 1,870	8%	152 / 1,854	8%	106 / 1,308	8%	102 / 1,246	8%
9	177 / 1,813	10%	182 / 1,849	10%	134 / 1,308	10%	140 / 1,246	11%
10	271 / 1,902	14%	265 / 1,857	14%	210 / 1,307	16%	216 / 1,245	17%

Table 4

Sensitivity, specificity, and positive predictive value (PPV) for the highest decile of risk for the most parsimonious equations using the internal and external validation populations.

	Sensitivity	Specificity	PPV
Internal validation of Midwestern HMO			
Equation A	24%	90%	14%
Equation B	25%	91%	15%
Equation C	28%	92%	17%
Equation D	28%	91%	16%
External validation of Midwestern HMO			
Equation A	20%	91%	26%
Equation B	20%	91%	26%
Equation C	21%	92%	28%
Equation D	22%	92%	29%
Internal validation of West Coast integrated delivery system			
Equation A	20%	91%	26%
Equation B	21%	91%	27%
Equation C	22%	91%	30%
Equation D	23%	92%	32%
External validation of West Coast integrated delivery system			
Equation A	24%	91%	14%
Equation B	23%	91%	14%
Equation C	26%	91%	16%
Equation D	28%	91%	17%