Online Supplement

Part A: Diabetes Outcome Definition on Claims Dataset

Includes:

Part B: Selection of top Variables for Presenting in Results

Includes:

Part C: Supplementary Tables for Results

Includes:

Part A: Diabetes Outcome Definition on Claims Dataset

A.1. Introduction

Medical claims data have become increasingly available to the medical research community, and could be a valuable resource for clinical observational studies. However, since claims data can be both inaccurate and incomplete, any prediction or analysis using these data must first define a reliable means of specifying the outcome definition. Inaccuracies in claims data can arise from many sources: The specific codes recorded may be biased toward those with a higher enumeration, because diagnosis codes are primarily only used for billing purposes. Additionally, not all actions, medications, and laboratory tests are recorded in the data. In the sections to follow, we describe our approach to remove the biases and mislabeling from our outcome label definition.

A.2. Outcome Labels

A.2.1. Type 2 Diabetes Outcome on Unbiased and Complete Data

On noise-free data, any of these conditions would indicate a beneficiary has Type 2 diabetes: (a) diagnosis by a physician with ICD-9 code 250.xx, (b) two measured HbA1c values higher than or equal to 6.5%, or (c) a pharmacy record for a Type 2 diabetes medication. In eTable A.1, we describe how different criteria compare to having any one of these conditions met. We used Glimepiride, Glipizide, Glyburide, Chlorpropamide, Tolazamide, Tolbutamide, Pioglitazone, Rosiglitazone, Acarbose, Miglitol, Repaglinide, Nateglinide, Sitagliptin, Saxagliptin, Linagliptin, Alogliptin, Pramlintide, Exenatide, Liraglutide, Canagliflozin, and Insulin (any) as medications for Type 2 diabetes in this outcome definition.

Table A.1. Coverage of Different Diabetes Outcome Definitions on Claims Data

Condition	Percentage (Raw Count)*
Have 250.x diagnosis, or have been on diabetic medication, or have any HbA1c $\geq$ 6.5	100 % (25638)
Have been diagnosed 250.xx	89.9 % (23051)
Have been on diabetic medications	15.0 % (3869)
Have HbA1c values $\geq$ 6.5	20.9 % (5360)
Have 250.xx diagnosis on more than one distinct date	40.0 % (10278)
(Have 250.xx diagnosis, or have been on diabetic medication, or have any HbA1c $\geq$ 6.5) on more than one distinct date	44.02 % (11286)
(Have 250.xx diagnosis, or have been on diabetic medication, or have any HbA1c $\geq$ 6.5) on two dates separated by at least a week	41.1 % (10558)

* Statistics calculated using data from January 1st 2011 to January 1st 2013, for patients continuously enrolled during this period who did not have diabetes before January 1st 2011.

A.2.2. Type 2 Diabetes Outcome on Claims Data

To define the outcome label on claims data, we required more than one evidence of diabetes, either on distinct days or over a longer span of time. To select the best outcome definition among the available choices, however, we needed to evaluate them on an unbiased gold standard set, in which we know for sure if a beneficiary is diabetic or non-diabetic.

We first build an accurate gold standard set. Focusing on beneficiaries who have had multiple tests for HbA1c, beneficiaries who had more than one HbA1c value above or equal to 6.5% within two years are considered truly diabetic.[1] If the beneficiary has only one instance of high HbA1c value, but is placed on diabetes medication, the beneficiary is also assumed to be truly diabetic. Conversely, if the beneficiary had two or more HbA1c tests performed and all of the results were below 6.5%, the beneficiary was not on diabetes medications, and the beneficiary did not have a diagnosis code for diabetes, we assumed them to be truly free from diabetes. eTable A.2 reports the number of beneficiaries satisfying each criteria.

Table A.2. Coverage of Different HbA1c Measurements on Original Gold Standard Set

Condition	Count (Percentage of those with 2 lab tests)
Total population which didn't have diabetes by 2011, and were enrolled between 2011 and 2013	473552
Have at least 2 HbA1c lab tests between 2011 and 2013	13975
Have absolute positive label between 2011 and 2013 (two HbA1c $\geq$ 6.5, or (one HbA1c $\geq$ 6.5 and on diabetes medication))	1473 (10.5%)
Have absolute negative label between 2011 and 2013 (at least two HbA1c tests done, but no HbA1c < 6.5, not on diabetes medication, and never diagnosed with 250.xx)	11123 (79.7%)
Have absolute positive or absolute negative labels between 2011 and 2013	12596 (90.2%)

However, these sets of the beneficiaries that are confirmed to either have diabetes or to be free from diabetes is not representative of the entire dataset. For example, the set of beneficiaries that have had two or more HbA1c tests may be biased by a variety of factors. Thus, we cannot select our outcome label definition by the performance of the label on this set as is.

Zubizarreta[2] showed that one can use mixed integer programming to select an optimal subset of individuals for which we know the labels, which match the entire population on several statistical measures. The statistical measures can be specified, and can include means, variances, and covariances of all beneficiary features.

For our purposes, we use age, gender, how long the beneficiary has been enrolled (measured by year), hypertension, hypercholesterolemia, and cardiovascular disease as the features, and select a subset of the beneficiaries with two HbA1c values, who have the same marginal and joint distribution of these features as the general population. Typically the matching is performed on the marginal distributions or pairwise joint distributions only. However, due to the large availability of data, we were able to match the joint distribution of these features as well, which gives a much closer distribution on the features than Zubizarreta's method.[2]

Our algorithm is as follows. We iteratively sampled without replacement from the joint distribution of the general population. For each sample, we selected a corresponding member from the gold standard whose features match exactly (age tolerance within ± 1 year). If there were multiple matches, we selected one uniformly from the available gold standard cases.

As the sampling continues, the likelihood of finding a match on features in the original gold standard set decreases. Hence, our algorithm stopped when we reached a mismatch rate of 8%, corresponding to the number of samples drawn from the original cohort which did not have a match in the gold standard set. The 8% mismatch rate was achieved when sampling 3000 individuals from original cohort. eTable A.3 shows the resulting distribution of features on the original gold standard (combining absolutely positive and absolutely negative sets), the unbiased gold standard, and the original general population.

Table A.3. Comparing Unbiased Gold Standard Label Set to Total Population*

Condition	Total Population*	Original Gold Label Set	Unbiased Gold Label Set
Size	473552	12596	2782
Average age (STD)	44.12(16.0)	55.19(13.8)	45.93(15.36)
Female ratio	53%	54%	56%
Average length of data in years (STD)	4.0(2.1)	4.3(2.1)	3.9(2.1)
Essential Hypertension ratio - ICD9 401	25.2%	55.1%	27.78%
Has Pure Hypercholesterolemia - ICD9 272.0	16.4%	33.7%	22.8%
Has Hypertensive Heart Disease - ICD9 402	2.5%	8.3%	5.1%

* Statistics calculated for beneficiaries with continuous enrollment from January 1st 2011 to January 1st 2013 for this set of experiments, who did not have diabetes before January 1st 2011.

A.3. Evaluation of Accuracy of Diabetes Outcome using Unbiased Gold Standard Set

We now have the gold standard set, in which we have identified true positive and true negative cases that are representative of the total population. Using this, we compare the accuracy of the following outcome definitions:

1. **Diabetes$_1$** = Any of the three: ICD-9 occurrence of 250.xx anywhere in the records, or any diabetes medicine, or any HbA1c value ≥ 6.5%
2. **Diabetes$_2$** = Any of the three: ICD-9 occurrence of 250.xx only in an outpatient clinical encounter or inpatient diagnosis (so long as there is no lab test order in that visit), or any diabetes medicine utilization (excluding Metformin), or any A1c value ≥ 6.5%
3. **Diabetes$_3$** = Observing definition Diabetes$_2$, for at least two distinct dates
4. **Diabetes$_4$** = Observing definition Diabetes$_2$, for at least two distinct dates separated by a week
5. **Diabetes$_5$** = Observing definition Diabetes$_2$, for at least two distinct dates separated by a month
6. **Diabetes$_6$** = Observing definition Diabetes$_2$, for at least two distinct dates separated by three months

We compute the Specificity, Sensitivity, Kappa, Precision, and F1-Harmonic mean of each of these outcome labels using the unbiased gold standard set. eTable A.4 shows these quality measures for each outcome label definition.

Table A.4. Evaluation of Different Outcome Definitions on Unbiased Gold Standard Set

Definition	Total Positive in Original Cohort (Percentage Among Full Cohort)	Sensitivity	Specificity	Kappa	Precision	F1
Diabetes$_1$	19050 (4.0%)	0.81	1	0.99	0.32	0.48
Diabetes$_2$	16651 (3.5%)	0.86	1	0.99	0.36	0.53
Diabetes$_3$	7448 (1.5%)	0.96	1	0.98	0.53	0.7
Diabetes$_4$	6978 (1.4%)	0.96	1	0.98	0.54	0.7
Diabetes$_5$	6453 (1.3%)	0.96	0.98	0.98	0.55	0.7
Diabetes$_6$	5726 (1.2%)	0.96	0.94	0.98	0.56	0.7

As is evident from eTable A.4, there is a trade-off between specificity and sensitivity. The choice of outcome measure depends on how the predictive models will be used: If the purpose of risk assessment is to identify a certain number of beneficiaries at highest risk, so as to perform an intervention on them, then a definition with high specificity should be employed, so that model predictions have highest accuracy for the most severe cases who will undergo intervention first. On the other hand, if a low cost intervention is

applied to all beneficiaries with slightest risk of developing diabetes, model fitting should be done with outcome labels with highest sensitivity.

We anticipate that the immediate uses of our population-level risk stratification models will be to perform interventions in a prioritized fashion, targeting the most severe cases first. Hence, we use **Diabetes$_3$** as our outcome definition, which has the highest-possible specificity of 1 and a sensitivity of 0.96 on the unbiased gold standard label set, and compared to Diabetes$_4$ has broader coverage without sacrificing specificity.

References

1. American Diabetes A. Standards of medical care in diabetes—2013. *Diabetes care.* 2013;36(Suppl 1):S11.
2. Zubizarreta JR. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association.* 2012;107(500):1360-1371.

Part B: Selection of top Variables for Presenting in Results

B.1. Selection of top Variables for Presenting in Results

L1-penalty guides the model to select a minimal set of variables that explain the outcome with highest accuracy. As we increase the L1-penalty or regularization, we are left with fewer but more predictive variables, i.e. features whose removal from the system would have significantly reduced the prediction accuracy.

Our best system selected via cross-validation has more than a hundred features as predictive. The additional features are important because many of these are surrogates for features that would otherwise be missing for many beneficiaries. In order to show top few features in Table 3 and 4 (in the main text) and Tables S5 (in Supplementary part C), we ran our system with a higher L1-penalty than selected with cross validation. This guided the system to select a much smaller set of representative features that is easier to visualize and interpret. This selection was only for the purpose of showing the 'top' variables in Tables 3, 4 and S5, and does not affect odds-ratio calculations, or final system.

Part C: Supplementary Tables

Table S1. Lab variables predictive of Type 2 diabetes confirmed onset between 2009 and 2011 (Gap=0), using patient data through Dec 31st 2008, sorted by the odds ratio of their prediction for Type 2 diabetes. Shown here are the variables with the highest magnitude of beta coefficient among all lab variables. There are a total of 226 lab variables predictive of future diabetes in this model.

Variable evaluation period*	Lab Variable Description	Number with diabetes	Number without diabetes	Odds ratio (95% CI)	OR for 18<age<40	OR for 40<age<65	OR for 65<age	p-value of OR
Past 6 months	loinc-4548-4 –Hemoglobin A1c/Hemoglobin.total -high	760	2720	11.6 (10.7 12.6)	30.89 (18.9 50.34)	10.77 (9.66 12.02)	5.26 (4.63 5.99)	<.001
Past 2 years	loinc-4548-4 –Hemoglobin A1c/Hemoglobin.total -high	1845	8708	9.28 (8.81 9.78)	23.01 (16.8 31.40)	8.42 (7.85 9.03)	4.34 (4.00 4.72)	<.001
Entire history	loinc-4548-4 –Hemoglobin A1c/Hemoglobin.total -high	2361	14488	7.30 (6.97 7.65)	11.17 (8.48 14.70)	6.85 (6.45 7.28)	3.44 (3.20 3.71)	<.001
Past 2 years	loinc-4548-4 –Hemoglobin A1c/Hemoglobin.total -increasing	639	3871	6.81 (6.25 7.41)	10.20 (6.05 17.20)	6.09 (5.45 6.81)	3.45 (3.01 3.96)	<.001
Entire history	loinc-27353-2 - Estimated average glucose -request for test	115	821	5.64 (4.64 6.86)	8.89 (3.91 20.20)	4.83 (3.77 6.18)	3.81 (2.68 5.42)	<.001
Past 6 months	loinc-2345-7 -Glucose -high	1485	11933	5.32 (5.03 5.63)	13.63 (10.21 18.20)	4.32 (4.01 4.66)	2.90 (2.66 3.17)	<.001
Past 6 months	loinc-4548-4 - Hemoglobin A1c/Hemoglobin.total -request for test	1784	16029	4.81 (4.57 5.07)	7.69 (6.09 9.70)	3.96 (3.70 4.23)	3.07 (2.82 3.34)	<.001
Past 2 years	loinc-2345-7 -Glucose -high	5274	58736	4.58 (4.43 4.73)	9.42 (7.90 11.24)	3.68 (3.52 3.84)	2.42 (2.29 2.56)	<.001
Entire history	loinc-2345-7 -Glucose -high	6936	89040	4.31 (4.18 4.44)	7.60 (6.51 8.87)	3.45 (3.32 3.59)	2.24 (2.12 2.36)	<.001
Past 2 years	loinc-4548-4 - Hemoglobin A1c/Hemoglobin.total -request for test	3908	45519	4.06 (3.92 4.21)	5.90 (5.03 6.91)	3.41 (3.25 3.57)	2.56 (2.40 2.73)	<.001
Entire history	loinc-2484-4 -Insulin-like growth factor-I -request for test	25	266	3.77 (2.50 5.68)	16.11 (6.98 37.16)	3.27 (1.96 5.47)	1.81 (0.55 5.90)	<.001
Entire history	loinc-4548-4 - Hemoglobin A1c/Hemoglobin.total -request for test	4995	66773	3.70 (3.58 3.82)	5.42 (4.70 6.25)	3.11 (2.98 3.25)	2.25 (2.12 2.38)	<.001
Past 6 months	loinc-2085-9 - Cholesterol.in HDL -increasing	349	4129	3.43 (3.07 3.83)	5.76 (2.71 12.25)	2.80 (2.40 3.26)	1.89 (1.60 2.22)	<.001
Entire history	loinc-2571-8 - Triglyceride -high	6056	106818	2.85 (2.77 2.94)	3.92 (3.37 4.55)	2.29 (2.20 2.38)	1.64 (1.55 1.73)	<.001
Entire history	loinc-6690-2 - Leukocytes -fluctuates	281	4010	2.84 (2.51 3.20)	3.30 (1.05 10.35)	2.27 (1.85 2.79)	1.51 (1.30 1.76)	<.001
Entire	loinc-13458-5 -	814	12055	2.78	4.31	2.25	1.64	<.001

history	Cholesterol.in VLDL - increasing			(2.59 2.99)	(2.73 6.81)	(2.05 2.47)	(1.45 1.84)	
Entire history	loinc-18262-6 - Cholesterol.in LDL - increasing	138	1998	2.78 (2.34 3.31)	6.98 (2.57 18.97)	2.19 (1.74 2.76)	1.60 (1.22 2.10)	<.001
Past 2 years	loinc-18262-6 - Cholesterol.in LDL - request for test	525	8504	2.52 (2.30 2.75)	4.32 (2.76 6.75)	1.98 (1.77 2.22)	1.71 (1.47 1.98)	<.001
Entire history	loinc-3094-0 -Urea nitrogen -fluctuates	1200	20178	2.48 (2.33 2.63)	3.10 (1.70 5.63)	1.97 (1.80 2.17)	1.29 (1.19 1.40)	<.001
Entire history	loinc-9830-1 - Cholesterol.total/Cholesterol.in HDL -high	3114	56032	2.46 (2.37 2.56)	4.12 (3.41 4.99)	2.04 (1.94 2.14)	1.47 (1.37 1.58)	<.001
Past 2 years	loinc-33914-3 – Glomerular fil. rate/1.73 sq M.pred -increasing	194	3284	2.38 (2.06 2.75)	-**	2.12 (1.56 2.89)	1.16 (0.99 1.37)	<.001
Entire history	loinc-2823-3 -Potassium -fluctuates	1170	20681	2.35 (2.21 2.50)	2.84 (1.56 5.17)	1.91 (1.74 2.09)	1.21 (1.12 1.32)	<.001
Entire history	loinc-6690-2 - Leukocytes -high	479	8412	2.31 (2.11 2.54)	3.50 (2.33 5.26)	1.99 (1.76 2.24)	1.51 (1.28 1.77)	<.001
Entire history	loinc-1742-6 -Alanine aminotransferase -high	1208	22205	2.26 (2.13 2.40)	3.49 (2.74 4.46)	2.00 (1.86 2.15)	1.53 (1.37 1.72)	<.001
Past 2 years	loinc-788-0 -Erythrocyte distribution width - decreasing	753	13723	2.25 (2.09 2.42)	3.01 (1.83 4.94)	1.74 (1.56 1.94)	1.33 (1.19 1.48)	<.001
Entire history	loinc-786-4 -Erythrocyte mean corpuscular hemo-globin concentration-low	230	4132	2.25 (1.97 2.57)	6.12 (3.51 10.66)	1.65 (1.38 1.96)	1.63 (1.31 2.04)	<.001
Past 2 years	loinc-4544-3 - Hematocrit -decreasing	1000	18445	2.24 (2.10 2.39)	2.46 (1.52 3.99)	1.80 (1.64 1.97)	1.27 (1.15 1.39)	<.001
Entire history	loinc-711-2 -Eosinophils -high	283	5137	2.23 (1.97 2.51)	2.98 (1.64 5.42)	1.80 (1.52 2.13)	1.45 (1.21 1.74)	<.001
Entire history	loinc-5902-2 - Coagulation tissue factor induced -request for test	656	12068	2.22 (2.05 2.40)	2.41 (1.36 4.28)	1.70 (1.51 1.93)	1.31 (1.17 1.46)	<.001
Past 2 years	loinc-713-8 – Eosinophils/100 leukocytes -increasing	676	12645	2.18 (2.02 2.36)	2.81 (1.58 4.98)	1.64 (1.46 1.84)	1.31 (1.17 1.47)	<.001
Entire history	loinc-13457-7 - Cholesterol.in LDL - request for test	11037	296992	2.14 (2.08 2.21)	2.58 (2.31 2.88)	1.49 (1.44 1.55)	1.32 (1.25 1.40)	<.001
Past 2 years	loinc-770-8 - Neutrophils/100 leukocytes -decreasing	774	15189	2.09 (1.94 2.25)	3.80 (2.43 5.93)	1.60 (1.44 1.78)	1.21 (1.09 1.35)	<.001
Past 2 years	loinc-19113-0 -IgE - request for test	38	744	2.05 (1.48 2.84)	1.96 (0.49 7.92)	1.90 (1.27 2.85)	1.89 (1.02 3.51)	<.001
Entire history	loinc-785-6 -Erythrocyte mean corpuscular hemoglobin -low	585	11746	2.03 (1.86 2.21)	3.90 (2.81 5.43)	1.68 (1.51 1.86)	1.50 (1.28 1.77)	<.001
Entire history	loinc-3084-1 -Urate - decreasing	111	2196	2.03 (1.68 2.46)	-**	1.61 (1.23 2.10)	1.20 (0.91 1.59)	<.001
Entire history	loinc-2086-7 –Choleste-rol.in HDL -decreasing	758	15736	1.97 (1.83 2.12)	3.88 (2.36 6.38)	1.56 (1.40 1.74)	1.06 (0.95 1.18)	<.001

*Entire history refers to our current setting and cohort, which is limited to max 4 years before 2009.
** If the age group with the predictive variable has zero size, no OR is reported

Table S2. Disease diagnosis history variables predictive of Type 2 diabetes confirmed onset between 2009 and 2011 (Gap=0), using patient data through Dec 31st 2008, sorted by the odds ratio of their prediction for Type 2 diabetes. Shown here are the variables with the highest magnitude of beta coefficient among the full ICD9 variable set. There are a total of 392 diagnosed conditions predictive of future diabetes in this model.

Disease History Variable Description*	Number with diabetes	Number without diabetes	Odds ratio (95% CI)	OR for 18<age<40	OR for 40<age<65	OR for 65<age	p-value of OR
ICD9 790.22 Impaired glucose tolerance test (oral)	265	1915	5.61 (4.93 6.38)	17.73 (11.1 28.13)	5.35 (4.51 6.33)	2.49 (1.99 3.11)	<.001
ICD9 271.9 Unspecified disorder of carbohydrate transport and metabolism	123	937	5.29 (4.38 6.39)	4.56 (1.12 18.57)	4.42 (3.43 5.69)	3.14 (2.35 4.20)	<.001
ICD9 790.29 Abnormal glucose	1198	10099	5.00 (4.70 5.32)	10.64 (7.89 14.35)	4.31 (3.98 4.68)	2.64 (2.39 2.92)	<.001
ICD9 790.21 Impaired fasting glucose	1285	11521	4.72 (4.45 5.01)	9.82 (6.69 14.41)	4.04 (3.74 4.37)	2.38 (2.16 2.62)	<.001
ICD9 271.3 Intestinal disaccharidase deficiencies and disaccharide malabsorption	397	3736	4.33 (3.90 4.80)	4.36 (2.56 7.43)	3.84 (3.34 4.41)	2.78 (2.35 3.30)	<.001
ICD9 405.11 Benign renovascular hypertension	29	276	4.22 (2.88 6.18)	-**	4.16 (2.43 7.11)	2.23 (1.28 3.88)	<.001
ICD9 277.7 Dysmetabolic syndrome x	513	5211	4.03 (3.67 4.41)	9.05 (6.48 12.64)	3.57 (3.19 3.99)	2.33 (1.94 2.79)	<.001
ICD9 428.22 Chronic systolic heart failure	142	1427	4.01 (3.37 4.77)	-**	3.87 (2.84 5.28)	1.94 (1.57 2.39)	<.001
ICD9 577.2 Cyst and pseudocyst of pancreas	48	502	3.84 (2.85 5.16)	5.95 (0.81 43.65)	3.61 (2.39 5.46)	1.83 (1.18 2.84)	<.001
ICD9 571.8 chronic nonalcoholic liver disease	619	6845	3.71 (3.41 4.03)	7.46 (5.22 10.66)	3.32 (3.01 3.66)	2.00 (1.68 2.39)	<.001
ICD9 571.5 Cirrhosis of liver without mention of alcohol	123	1373	3.61 (3.00 4.34)	12.76 (4.62 35.19)	3.27 (2.63 4.07)	1.49 (1.03 2.17)	<.001
ICD9 410.91 Acute myocardial infarction of unspecified site, initial episode of care	93	1037	3.61 (2.92 4.46)	-**	3.31 (2.46 4.45)	1.67 (1.23 2.27)	<.001
ICD9 401.9 Unspecified essential hypertension	8755	149273	3.47 (3.37 3.57)	5.16 (4.46 5.97)	2.69 (2.59 2.79)	1.64 (1.55 1.73)	<.001
ICD9 584.9 Acute kidney failure, unspecified	297	3537	3.40 (3.02 3.83)	4.22 (1.74 10.27)	2.95 (2.42 3.60)	1.78 (1.53 2.07)	<.001
ICD9 272.5 Lipoprotein deficiencies	129	1555	3.34 (2.79 4.00)	5.82 (2.39 14.20)	2.86 (2.29 3.57)	1.96 (1.42 2.72)	<.001
ICD9 428.0 Congestive heart failure, unspecified	849	10617	3.31 (3.08 3.55)	2.10 (0.78 5.63)	2.93 (2.60 3.31)	1.66 (1.52 1.82)	<.001
ICD9 401.1 Benign essential hypertension	9047	164088	3.28 (3.18 3.37)	4.31 (3.69 5.04)	2.44 (2.35 2.53)	1.59 (1.51 1.68)	<.001
ICD9 414.01 Coronary atherosclerosis of native coronary artery	2265	31250	3.16 (3.02 3.30)	3.61 (1.93 6.76)	2.46 (2.29 2.64)	1.66 (1.56 1.77)	<.001
ICD9 410.90 Acute	184	2353	3.15	0.00	2.44	1.70	<.001

myocardial infarction of unspecified site, episode of care unspecified			(2.71 3.67)	(0.00 nan)	(1.93 3.09)	(1.39 2.07)	
ICD9 789.1 Hepatomegaly	252	3260	3.13 (2.75 3.56)	5.62 (3.41 9.27)	2.63 (2.24 3.08)	2.37 (1.85 3.05)	<.001
ICD9 V458.1 Aortocoronary bypass status	383	5014	3.10 (2.79 3.45)	-**	2.79 (2.25 3.46)	1.52 (1.34 1.71)	<.001
ICD9 278.01 Morbid obesity	871	11798	3.05 (2.84 3.27)	6.17 (4.92 7.74)	2.81 (2.59 3.06)	2.19 (1.86 2.59)	<.001
ICD9 253.3 Pituitary dwarfism	25	331	3.03 (2.02 4.55)	14.70 (8.81 24.52)	1.90 (0.83 4.32)	2.71 (0.81 9.07)	<.001
ICD9 429.3 Cardiomegaly	876	12098	2.99 (2.79 3.21)	4.12 (2.37 7.15)	2.55 (2.29 2.84)	1.54 (1.40 1.70)	<.001
ICD9 402.10 Benign hypertensive heart disease without heart failure	1015	14195	2.97 (2.78 3.17)	5.02 (2.89 8.74)	2.33 (2.12 2.56)	1.62 (1.47 1.78)	<.001
ICD9 274.9 Gout	711	9844	2.97 (2.75 3.21)	3.67 (1.82 7.40)	2.36 (2.12 2.63)	1.65 (1.47 1.85)	<.001
ICD9 414.00 Coronary atherosclerosis of unspecified type of vessel, native or graft	1846	26812	2.95 (2.80 3.10)	3.38 (1.74 6.54)	2.30 (2.12 2.50)	1.54 (1.44 1.64)	<.001
ICD9 278.00 Obesity	2494	37342	2.93 (2.80 3.06)	4.56 (3.92 5.30)	2.92 (2.77 3.08)	1.89 (1.72 2.07)	<.001
ICD9 277.9 Metabolism disorder	119	1648	2.91 (2.41 3.50)	4.80 (2.26 10.19)	2.45 (1.94 3.11)	2.08 (1.49 2.91)	<.001
ICD9 411.1 Intermediate coronary syndrome	508	7143	2.90 (2.65 3.18)	7.31 (3.23 16.56)	2.30 (2.02 2.62)	1.54 (1.35 1.75)	<.001
ICD9 327.23 Obstructive sleep apnea	1178	17302	2.84 (2.67 3.02)	4.11 (3.07 5.50)	2.48 (2.30 2.66)	1.81 (1.60 2.05)	<.001
ICD9 780.57 Sleep apnea	919	13478	2.82 (2.63 3.02)	4.45 (3.21 6.17)	2.41 (2.22 2.61)	1.84 (1.61 2.11)	<.001
ICD9 780.53 Hypersomnia with sleep apnea, unspecified	1138	16965	2.79 (2.63 2.97)	4.15 (3.04 5.67)	2.38 (2.21 2.56)	1.83 (1.62 2.08)	<.001
ICD9 429.9 Heart disease	207	3030	2.76 (2.39 3.18)	1.93 (0.48 7.79)	2.57 (2.09 3.17)	1.38 (1.13 1.68)	<.001

*ICD9 variables are evaluated over the entire past history, which is currently limited to 4 years before 2009.
** If the age group with the predictive variable has zero size, no OR is reported

Table S3. Lab variables predictive of Type 2 diabetes confirmed onset between 2010 and 2012, using patient data through Dec 31st 2008 (Gap=1), sorted by the odds ratio of their prediction for Type 2 diabetes. Shown here are the variables with the highest magnitude of beta coefficient among all lab variables. There are a total of 175 lab variables predictive of future diabetes in this model.

Variable evaluatio period*	Lab Variable Description	Number with diabetes	Number without diabetes	Odds ratio (95% CI)	OR for 18<age<40	OR for 40<age<65	OR for 65<age	p-value of OR
Past 2 years	loinc-4548-4 –Hemo-globin A1c/Hemo-globin.total -high	981	7308	7.06 (6.59 7.57)	14.94 (9.74 22.91)	6.48 (5.92 7.09)	3.35 (3.00 3.74)	<.001
Entire history	loinc-4548-4 –Hemo-globin A1c/Hemo-globin.total -high	1323	12344	5.75 (5.42 6.10)	7.98 (5.58 11.41)	5.46 (5.05 5.90)	2.74 (2.49 3.02)	<.001
Entire history	loinc-2484-4 -Insulin-like growth factor-I -request for test	25	224	5.52 (3.65 8.35)	22.93 (10.4 50.22)	4.17 (2.40 7.22)	4.25 (1.47 12.25)	<.001
Past 6 months	loinc-2345-7 -Glucose -high	884	10166	4.52 (4.21 4.85)	7.15 (4.61 11.09)	3.57 (3.25 3.93)	2.72 (2.43 3.03)	<.001
Past 2 years	loinc-2345-7 -Glucose -high	3389	50745	4.05 (3.89 4.21)	7.31 (5.88 9.10)	3.24 (3.07 3.41)	2.25 (2.10 2.40)	<.001
Past 2 years	loinc-13458-5 -Cholesterol.in VLDL -increasing	467	6097	3.88 (3.53 4.27)	6.38 (3.58 11.38)	2.86 (2.51 3.26)	2.61 (2.25 3.03)	<.001
Entire history	loinc-2345-7 -Glucose -high	4551	77532	3.83 (3.70 3.97)	6.16 (5.12 7.42)	3.05 (2.91 3.20)	2.10 (1.96 2.23)	<.001
Past 2 years	loinc-4548-4 -Hemoglobin A1c/Hemoglobin.total -request for test	2389	39347	3.42 (3.27 3.58)	5.11 (4.23 6.17)	2.90 (2.74 3.07)	2.14 (1.97 2.32)	<.001
Past 6 months	loinc-13458-5 -Cholesterol.in VLDL -request for test	822	13236	3.20 (2.98 3.44)	4.06 (2.83 5.83)	2.54 (2.31 2.79)	2.32 (2.05 2.62)	<.001
Entire history	loinc-4548-4 -Hemoglobin A1c/Hemoglobin.total -request for test	3111	58061	3.13 (3.00 3.26)	4.63 (3.91 5.47)	2.63 (2.49 2.77)	1.94 (1.81 2.09)	<.001
Entire history	loinc-2823-3 -Potassium -low	199	3841	2.58 (2.24 2.98)	2.34 (0.97 5.66)	2.47 (2.04 2.98)	1.36 (1.09 1.71)	<.001
Entire history	loinc-9830-1 -Cholesterol.total/Cholesterol.in HDL -high	2082	49026	2.29 (2.19 2.40)	4.00 (3.22 4.97)	1.87 (1.76 1.98)	1.42 (1.30 1.55)	<.001
Entire history	loinc-786-4 -Erythrocyte mean corpuscular hemoglobin concentration -low	162	3575	2.25 (1.92 2.64)	7.41 (4.15 13.22)	1.52 (1.23 1.88)	1.82 (1.41 2.35)	<.001
Entire history	loinc-13458-5 -Cholesterol.in VLDL -request for test	2277	55592	2.23 (2.13 2.33)	2.43 (1.96 3.01)	1.80 (1.70 1.91)	1.67 (1.54 1.81)	<.001
Entire history	loinc-711-2 -Eosinophils -high	190	4489	2.11 (1.82 2.44)	3.04 (1.57 5.89)	1.62 (1.32 1.99)	1.49 (1.20 1.86)	<.001
Entire history	loinc-1558-6 -Glucose^post CFst -request for test	191	4545	2.09 (1.81 2.42)	3.50 (2.27 5.41)	1.80 (1.49 2.18)	1.95 (1.49 2.54)	<.001

Entire history	loinc-33914-3 -Glomerular filtration rate/1.73 sq M.predicted -low	743	18216	2.07 (1.92 2.24)	1.10 (0.15 7.88)	1.61 (1.40 1.84)	1.09 (0.99 1.19)	<.001
Past 6 months	loinc-2349-9 -Glucose -request for test	257	6186	2.07 (1.83 2.35)	1.68 (0.92 3.05)	1.81 (1.54 2.12)	1.82 (1.47 2.27)	<.001
Entire history	loinc-1742-6 -Alanine aminotransferase -high	775	19358	2.04 (1.89 2.19)	3.18 (2.39 4.23)	1.78 (1.63 1.95)	1.42 (1.23 1.64)	<.001
Entire history	loinc-13457-7 -Cholesterol.in LDL -increasing	2338	67196	1.87 (1.78 1.95)	2.47 (1.88 3.23)	1.37 (1.29 1.45)	1.19 (1.10 1.28)	<.001
Entire history	loinc-2857-1 -Prostate specific Ag -request for test	1830	52668	1.83 (1.74 1.92)	2.66 (1.46 4.84)	1.34 (1.26 1.43)	1.08 (0.99 1.17)	<.001
Past 2 years	loinc-2093-3 -Cholesterol -high	3195	96919	1.82 (1.75 1.89)	3.01 (2.49 3.63)	1.40 (1.33 1.47)	1.11 (1.03 1.19)	<.001
Past 2 years	loinc-9830-1 -Cholesterol.total/Cholesterol.in HDL -decreasing	1454	42342	1.78 (1.68 1.88)	2.69 (1.82 3.98)	1.33 (1.23 1.43)	1.04 (0.95 1.13)	<.001
Past 2 years	loinc-630-4 -Bacteria identified -request for test	230	6590	1.74 (1.52 1.98)	3.47 (2.08 5.81)	1.44 (1.20 1.72)	1.08 (0.87 1.35)	<.001
Entire history	loinc-1971-1 -Bilirubin.non-glucuronidated -decreasing	207	6119	1.68 (1.46 1.93)	1.48 (0.55 3.95)	1.50 (1.24 1.80)	0.93 (0.75 1.16)	<.001
Entire history	loinc-1759-0 -Albumin/Globulin -increasing	1347	41719	1.66 (1.57 1.76)	2.86 (2.02 4.06)	1.21 (1.12 1.31)	1.02 (0.94 1.12)	<.001
Entire history	loinc-3097-3 -Urea nitrogen/Creatinine -high	910	28625	1.61 (1.50 1.72)	2.22 (1.31 3.77)	1.13 (1.02 1.25)	0.97 (0.89 1.07)	<.001
Past 6 months	loinc-33914-3 -Glomerular filtration rate/1.73 sq M.predicted -request for test	2068	67751	1.60 (1.52 1.68)	2.39 (1.91 3.00)	1.19 (1.11 1.26)	1.07 (0.99 1.16)	<.001
Past 2 years	loinc-5811-5 -Specific gravity increasing	113	3520	1.59 (1.32 1.92)	-**	1.13 (0.86 1.49)	1.07 (0.83 1.39)	<.001
Entire history	loinc-5196-1 -Hepatitis B virus surface Ag -request for test	309	9843	1.56 (1.39 1.75)	2.68 (1.93 3.71)	1.53 (1.33 1.75)	1.35 (1.02 1.79)	<.001
Entire history	loinc-785-6 -Erythrocyte mean corpuscular hemoglobin -high	469	16857	1.39 (1.26 1.52)	2.50 (1.41 4.43)	0.96 (0.84 1.09)	0.94 (0.82 1.07)	<.001
Past 2 years	loinc-1975-2 -Bilirubin -high	160	6942	1.14 (0.97 1.34)	1.10 (0.46 2.66)	0.80 (0.64 1.00)	1.01 (0.80 1.28)	0.102
Entire history	loinc-10834-0 -Globulin -low	195	9109	1.06 (0.92 1.22)	0.61 (0.15 2.47)	0.75 (0.62 0.92)	0.79 (0.64 0.97)	0.434
Entire history	loinc-5060-9 -Borrelia burgdorferi Ab -request for test	160	8282	0.95 (0.82 1.12)	0.65 (0.24 1.73)	0.87 (0.73 1.05)	0.64 (0.46 0.89)	0.559
Past 2 years	loinc-5060-9 -Borrelia burgdorferi Ab -request for test	82	4940	0.82 (0.66 1.02)	0.72 (0.23 2.23)	0.75 (0.58 0.96)	0.60 (0.38 0.94)	0.074

*Entire history refers to our current setting and cohort, which is limited to max 4 years before 2009.
** If the age group with the predictive variable has zero size, no OR is reported

Table S4. Disease diagnosis history variables predictive of Type 2 diabetes confirmed onset between 2010 and 2012 (Gap=1), using patient data through Dec 31st 2008, sorted by the odds ratio of their prediction for Type 2 diabetes. Shown here are the variables with the highest magnitude of beta coefficient among the full ICD9 variable set. There are a total of 303 diagnosed condition variables predictive of future diabetes.

Disease History Variable Description*	Number with diabetes	Number without diabetes	Odds ratio (95% CI)	OR for 18<age<40	OR for 40<age<65	OR for 65<age	p-value of OR
ICD9 790.21 Impaired fasting glucose	800	9918	4.17 (3.87 4.49)	7.05 (4.27 11.65)	3.42 (3.10 3.77)	2.33 (2.07 2.62)	<.001
ICD9 790.22 Impaired glucose tolerance test (oral)	138	1661	4.14 (3.47 4.93)	9.48 (4.85 18.55)	3.86 (3.07 4.84)	2.04 (1.51 2.75)	<.001
ICD9 790.29 Abnormal glucose	690	8695	4.07 (3.76 4.41)	7.46 (5.05 11.00)	3.46 (3.12 3.84)	2.28 (2.01 2.60)	<.001
ICD9 571.8 Chronic nonalcoholic liver disease	422	5891	3.62 (3.27 4.00)	6.67 (4.38 10.14)	3.26 (2.90 3.67)	1.92 (1.54 2.40)	<.001
ICD9 253.3 Pituitary dwarfism	20	283	3.50 (2.22 5.50)	15.34 (8.70 27.04)	2.68 (1.17 6.11)	1.42 (0.19 10.60)	<.001
ICD9 327.23 Obstructive sleep apnea	867	14979	2.98 (2.78 3.20)	4.50 (3.30 6.15)	2.61 (2.40 2.84)	1.89 (1.63 2.19)	<.001
ICD9 799.89 Illness, unspecified	68	1153	2.92 (2.29 3.74)	3.73 (1.39 10.05)	2.81 (2.03 3.90)	1.74 (1.16 2.62)	<.001
ICD9 789.1 Hepatomegaly	159	2851	2.78 (2.36 3.26)	5.09 (2.86 9.06)	2.44 (2.01 2.96)	1.88 (1.34 2.64)	<.001
ICD9 790.6 Other abnormal blood chemistry	1588	33877	2.49 (2.36 2.62)	3.81 (2.99 4.86)	2.08 (1.94 2.23)	1.51 (1.38 1.65)	<.001
ICD9 V43.65 Knee joint replacement	269	5568	2.41 (2.13 2.73)	0.00 (0.00 nan)	2.31 (1.89 2.83)	1.16 (0.99 1.36)	<.001
ICD9 070.54 Chronic hepatitis C without mention of hepatic coma	147	3145	2.32 (1.97 2.74)	1.22 (0.17 8.75)	2.06 (1.72 2.46)	1.12 (0.70 1.80)	<.001
ICD9 573.3 Hepatitis	226	4893	2.30 (2.01 2.64)	3.62 (2.20 5.95)	2.18 (1.86 2.56)	1.34 (0.99 1.79)	<.001
ICD9 253.4 Other anterior pituitary disorders	52	1122	2.30 (1.74 3.03)	8.21 (4.04 16.69)	2.07 (1.51 2.84)	0.69 (0.22 2.19)	<.001
ICD9 726.73 Calcaneal spur	466	10686	2.20 (2.00 2.41)	4.92 (3.03 7.99)	1.71 (1.52 1.91)	1.53 (1.28 1.84)	<.001
ICD9 242.90 Thyrotoxicosis without mention of goiter or other cause, and without mention of thyrotoxic crisis or storm	417	9570	2.19 (1.98 2.42)	4.55 (3.21 6.46)	1.90 (1.67 2.15)	1.34 (1.12 1.61)	<.001
ICD9 786.9 Symptoms involving respiratory system and chest	216	4930	2.18 (1.90 2.50)	3.13 (1.76 5.55)	1.86 (1.55 2.23)	1.47 (1.17 1.85)	<.001
ICD9 280.1 Iron deficiency anemia secondary to inadequate dietary iron intake	189	4687	2.01 (1.73 2.32)	3.24 (1.97 5.33)	1.75 (1.45 2.12)	1.51 (1.16 1.96)	<.001
ICD9 648.81 Abnormal glucose tolerance of mother, delivered, with or	37	958	1.91 (1.38 2.65)	7.91 (5.04 12.41)	1.60 (0.99 2.60)	-**	<.001

without mention of antepartum condition							
ICD9 427.81 Sinoatrial node dysfunction	210	5799	1.80 (1.57 2.07)	1.88 (0.70 5.04)	1.33 (1.05 1.70)	1.08 (0.91 1.28)	<.001
ICD9 592.1 Calculus of ureter	239	6705	1.77 (1.56 2.02)	2.05 (1.13 3.72)	1.53 (1.31 1.79)	1.39 (1.07 1.80)	<.001
ICD9 424.0 Mitral valve disorder	1070	32632	1.67 (1.57 1.78)	1.81 (1.19 2.74)	1.16 (1.06 1.28)	1.13 (1.03 1.24)	<.001
ICD9 701.9 Hypertrophic and atrophic conditions of skin	402	12090	1.66 (1.50 1.84)	1.25 (0.67 2.34)	1.44 (1.28 1.62)	1.21 (0.99 1.48)	<.001
ICD9 256.4 Polycystic ovaries	109	3285	1.64 (1.36 1.99)	4.97 (3.74 6.60)	2.10 (1.61 2.74)	1.59 (0.21 12.01)	<.001
ICD9 727.05 Tenosynovitis of hand and wrist	155	4713	1.63 (1.39 1.92)	1.24 (0.56 2.78)	1.45 (1.19 1.78)	1.40 (1.04 1.87)	<.001
ICD9 719.51 Stiffness of joint, not elsewhere classified, shoulder region	58	1948	1.47 (1.13 1.91)	0.51 (0.07 3.60)	1.36 (0.99 1.87)	1.26 (0.77 2.05)	.0037
ICD9 783.1 Abnormal weight gain	522	17988	1.45 (1.33 1.59)	2.42 (1.83 3.19)	1.32 (1.18 1.46)	1.34 (1.10 1.63)	<.001
ICD9 461.1 Acute frontal sinusitis	191	6586	1.44 (1.25 1.66)	1.57 (0.93 2.67)	1.43 (1.20 1.69)	1.27 (0.91 1.76)	<.001
ICD9 7955 Nonspecific reaction to test for tuberculosis	125	4333	1.43 (1.20 1.71)	1.12 (0.56 2.24)	1.54 (1.25 1.90)	1.51 (1.00 2.27)	<.001
ICD9 64883 Abnormal glucose tolerance of mother, antepartum condition or complication	53	1872	1.40 (1.07 1.84)	4.79 (3.18 7.21)	1.42 (0.98 2.06)	-**	.015
ICD9 388.71 Otogenic pain	50	1794	1.38 (1.04 1.83)	2.25 (0.93 5.45)	1.21 (0.84 1.75)	1.15 (0.69 1.94)	.025
ICD9 611.1 Hypertrophy of breast	121	4400	1.36 (1.14 1.63)	2.15 (1.29 3.59)	1.35 (1.08 1.68)	1.28 (0.85 1.93)	.0008
ICD9 614.6 Pelvic peritoneal adhesions, female (postoperative) (postinfection)	75	2730	1.36 (1.08 1.71)	4.38 (2.66 7.22)	1.23 (0.94 1.61)	0.68 (0.25 1.85)	.008
ICD9 370.33 Keratoconjunctivitis sicca, not specified as Sjogren's	131	4896	1.33 (1.11 1.58)	1.31 (0.42 4.09)	0.81 (0.62 1.07)	1.07 (0.85 1.35)	.0015
ICD9 783.43 Short stature	21	944	1.10 (0.71 1.69)	4.61 (2.95 7.21)	0.79 (0.11 5.71)	-**	.66
ICD9 626.4 Irregular menstruation	404	23114	0.86 (0.78 0.95)	2.59 (2.12 3.16)	0.84 (0.75 0.95)	1.16 (0.54 2.47)	.003

*ICD9 variables are evaluated over the entire past history, which is currently limited to max 4 years before 2009.

** If the age group with the predictive variable has zero size, no OR is reported

Table S5. Top predictive variables for Type 2 diabetes onset within window 2011-2013 (Gap=2), using patient data through Dec 31st 2008. We use the methodology described in Supplement Part B to select these variables.

Variable Type	Variable evaluation period*	Variable Description	Number with diabetes	Number without diabetes	Odds ratio (95% CI)	OR for 18≤age<40	OR for 40≤age<65	OR for 65≤age	p-value of OR
Lab test	Entire history	Hemoglobin A1c / Hemoglobin.total - high (Loinc-4548-4)	610	10981	4.30 (3.95 4.68)	6.96 (4.33 11.17)	4.09 (3.67 4.56)	2.08 (1.80 2.40)	<.001
	Entire history	Glucose –high (loinc-2345-7)	2489	69907	3.27 (3.12 3.43)	4.74 (3.67 6.11)	2.61 (2.45 2.77)	1.90 (1.74 2.07)	<.001
	Entire history	Cholesterol.in HDL -low (loinc-2085-9)	1270	38738	2.64 (2.49 2.81)	3.72 (2.87 4.82)	2.20 (2.04 2.36)	1.81 (1.61 2.04)	<.001
	Entire history	Cholesterol.total/Cholesterol.in HDL - high (loinc-9830-1)	1317	44487	2.38 (2.24 2.52)	4.05 (3.10 5.28)	1.90 (1.76 2.04)	1.57 (1.40 1.76)	<.001
	Entire history	Hemoglobin A1c / Hemoglobin.total - request for test (Loinc-4548-4)	1486	52621	2.29 (2.16 2.42)	3.58 (2.85 4.49)	1.93 (1.80 2.07)	1.43 (1.29 1.59)	<.001
ICD9 diagnosis history	Entire history	Unspecified essential hypertension (icd9 401.9)	4954	180103	3.42 (3.28 3.58)	4.23 (3.54 5.06)	2.58 (2.45 2.72)	1.69 (1.52 1.87)	<.001
	Past 2 years	Unspecified essential hypertension (icd9 401.9)	4526	158769	3.32 (3.18 3.47)	4.64 (3.85 5.59)	2.58 (2.45 2.72)	1.56 (1.41 1.72)	<.001
	Entire history	Has obesity and other hyper-alimentation (icd9 278)	1316	37694	2.84 (2.67 3.01)	4.60 (3.80 5.57)	2.77 (2.58 2.97)	1.74 (1.50 2.02)	<.001
	Entire history	Hyperlipidemia (icd9 272.4)	3464	148333	2.19 (2.10 2.29)	2.79 (2.26 3.44)	1.63 (1.54 1.72)	1.21 (1.11 1.32)	<.001
Healthcare utilization	Past 2 years	Specialty visit: cardiology	2487	102919	2.08 (1.99 2.18)	2.09 (1.64 2.66)	1.56 (1.47 1.66)	1.42 (1.30 1.55)	<.001
	Entire history	Procedure group: Routine chest X ray	2919	137152	1.85 (1.77 1.93)	1.96 (1.64 2.35)	1.49 (1.41 1.58)	1.34 (1.23 1.46)	<.001
	Entire history	Service place: Emergency Room Hospital	3641	223629	1.33 (1.28 1.39)	1.35 (1.16 1.58)	1.39 (1.31 1.47)	1.43 (1.31 1.56)	<.001
	Entire history	Screening mammogram (icd9 V76.12)	2292	142643	1.24 (1.18 1.30)	2.02 (0.90 4.53)	0.82 (0.77 0.87)	0.76 (0.69 0.83)	<.001
	Entire history	Service place -Office	8272	605650	0.95 (0.83 1.08)	0.63 (0.44 0.91)	0.99 (0.85 1.17)	0.72 (0.53 0.99)	<.001
	Past 6 months	Pharmacy coverage indicator =Yes	4189	316331	0.94 (0.90 0.98)	0.91 (0.78 1.05)	0.79 (0.75 0.83)	0.80 (0.72 0.88)	0.003

	Past 2 years	Vision coverage indicator = No	5102	382389	0.94 (0.90 0.98)	1.01 (0.87 1.18)	1.07 (1.01 1.13)	0.85 (0.78 0.93)	0.005
	Entire history	Procedure group: Prophylactic vaccinations and inoculations	3421	262533	0.92 (0.88 0.96)	0.53 (0.46 0.62)	1.06 (1.00 1.12)	0.95 (0.87 1.04)	<.001
	Entire history	Routine medical examination (icd9 V70.0)	2116	174721	0.85 (0.81 0.89)	1.21 (1.02 1.43)	0.76 (0.71 0.80)	0.70 (0.63 0.78)	<.001
	Entire history	Routine gynecological examination (icd9 V72.31)	2692	224702	0.82 (0.78 0.86)	1.44 (1.24 1.68)	0.68 (0.64 0.72)	0.79 (0.72 0.88)	<.001
	Entire history	Specialty visit: dermatology	1893	162854	0.81 (0.77 0.85)	0.70 (0.58 0.85)	0.73 (0.68 0.77)	0.79 (0.72 0.87)	<.001
	Entire history	Specialty visit: obstetrics-gynecology	2139	191164	0.76 (0.72 0.80)	1.41 (1.21 1.65)	0.65 (0.61 0.69)	0.80 (0.72 0.89)	<.001
	Past 2 years	Specialty visit obstetrics-gynecology	1873	173667	0.73 (0.69 0.77)	1.38 (1.18 1.61)	0.63 (0.59 0.67)	0.77 (0.68 0.86)	<.001
	Entire history	Specialty visit: pediatrics	187	72947	0.17 (0.15 0.20)	0.33 (0.27 0.41)	0.65 (0.52 0.81)	0.56 (0.30 1.04)	<.001
	Entire history	Routine child health exam (icd9 V202)	119	70898	0.11 (0.09 0.13)	0.31 (0.25 0.38)	0.44 (0.29 0.68)	0.68 (0.09 4.93)	<.001

*Entire history refers to our current setting and cohort, which is limited to max 4 years before 2009.