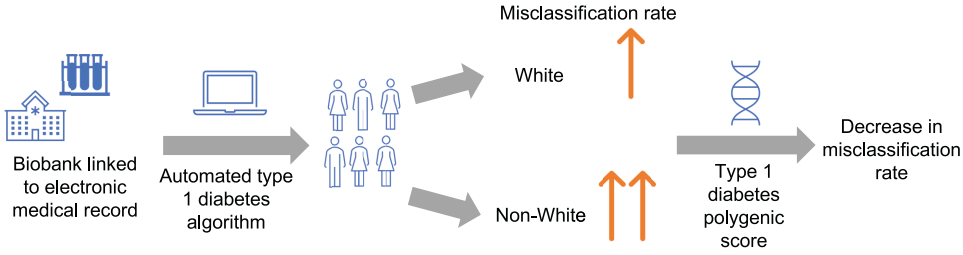
Diabetes Care
American Diabetes Association

Polygenic Scores Help Reduce Racial Disparities in Predictive Accuracy of Automated Type 1 Diabetes Classification Algorithms

Aaron J. Deutsch, Lauren Stalbow, Timothy D. Majarian, Josep M. Mercader, Alisa K. Manning, Jose C. Florez, Ruth J.F. Loos, and Miriam S. Udler

Diabetes Care 2023;46(4):794–800 | https://doi.org/10.2337/dc22-1833



ARTICLE HIGHLIGHTS

- Automated algorithms using electronic health record data can identify individuals with type 1 diabetes in large biobanks, but it is not known whether the accuracy of these algorithms differs according to self-reported race.
- This study shows that misclassification of type 1 diabetes is more likely in self-reported non-White participants than in self-reported White participants.
- Incorporating genetic information using type 1 diabetes polygenic scores can improve the accuracy of classification algorithms.
- These findings highlight a disparity in existing type 1 diabetes classification algorithms and propose a potential solution.

Polygenic Scores Help Reduce Racial Disparities in Predictive Accuracy of Automated Type 1 Diabetes Classification Algorithms

Aaron J. Deutsch,[1–3] Lauren Stalbow,[4] Timothy D. Majarian,[2] Josep M. Mercader,[1–3] Alisa K. Manning,[2,3,5] Jose C. Florez,[1–3] Ruth J.F. Loos,[4,6] and Miriam S. Udler[1–3]

Diabetes Care 2023;46:794–800 | https://doi.org/10.2337/dc22-1833

ORIGINAL ARTICLE

OBJECTIVE

Automated algorithms to identify individuals with type 1 diabetes using electronic health records are increasingly used in biomedical research. It is not known whether the accuracy of these algorithms differs by self-reported race. We investigated whether polygenic scores improve identification of individuals with type 1 diabetes.

RESEARCH DESIGN AND METHODS

We investigated two large hospital-based biobanks (Mass General Brigham [MGB] and Bio*Me*) and identified individuals with type 1 diabetes using an established automated algorithm. We performed medical record reviews to validate the diagnosis of type 1 diabetes. We implemented two published polygenic scores for type 1 diabetes (developed in individuals of European or African ancestry). We assessed the classification algorithm before and after incorporating polygenic scores.

RESULTS

The automated algorithm was more likely to incorrectly assign a diagnosis of type 1 diabetes in self-reported non-White individuals than in self-reported White individuals (odds ratio 3.45; 95% CI 1.54–7.69; *P* = 0.0026). After incorporating polygenic scores into the MGB Biobank, the positive predictive value of the type 1 diabetes algorithm increased from 70 to 97% for self-reported White individuals (meaning that 97% of those predicted to have type 1 diabetes indeed had type 1 diabetes) and from 53 to 100% for self-reported non-White individuals. Similar results were found in Bio*Me*.

CONCLUSIONS

Automated phenotyping algorithms may exacerbate health disparities because of an increased risk of misclassification of individuals from underrepresented populations. Polygenic scores may be used to improve the performance of phenotyping algorithms and potentially reduce this disparity.

Biobanks linked to electronic health records (EHRs) offer a wealth of clinical information, presenting opportunities for research in large numbers of individuals, as reported, for example, by the All of Us research program (1). However, the extraction of accurate phenotype information from EHR data can be challenging. Because

[1]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA
[2]Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA
[3]Department of Medicine, Harvard Medical School, Boston, MA
[4]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
[5]Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA
[6]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

Corresponding author: Miriam S. Udler, mudler@mgh.harvard.edu

Received 19 September 2022 and accepted 10 January 2023

This article contains supplementary material online at https://doi.org/10.2337/figshare.21893889.

This article is featured in podcasts available at diabetesjournals.org/care/pages/diabetes_care_on_air.

A.J.D. and L.S. contributed equally to this work.

T.D.M. is currently affiliated with Vertex Pharmaceuticals, Boston, MA.

diagnosis codes are primarily recorded for billing purposes, they may not accurately capture relevant phenotypes (2). Instead of relying solely on diagnosis codes, targeted strategies are necessary to extract phenotypes of interest from the vast array of EHR data (which include categories such as demographics, laboratory values, and medication prescriptions in addition to diagnosis codes) (3,4); still, incomplete EHR data present an additional challenge. Multiple algorithms have been developed and validated that use EHR data to identify individuals with type 1 and type 2 diabetes (5–11). However, it is unknown whether these algorithms accurately capture the intended phenotype across diverse populations in the medical system. Specifically, misclassification of disease is frequently present in studies based on EHR data, and sociodemographic factors such as race and ethnicity may play a role in misclassification (12). In this work, we assessed the accuracy of a type 1 diabetes classification algorithm across two large health systems in the U.S., with a focus on racial variation in misclassification rates.

Because EHR-derived type 1 diabetes definitions can lead to misclassification, we investigated whether the inclusion of genetic information could improve classification accuracy. Previous work has demonstrated that genetic information (through the use of type 1 diabetes polygenic scores) can identify individuals with type 1 diabetes with high accuracy (13,14). Here, we demonstrate that the incorporation of type 1 diabetes polygenic scores can enhance existing algorithm-based identification of individuals with type 1 diabetes and reduce the disparity in misclassification rates among racial groups.

RESEARCH DESIGN AND METHODS

A schematic flowchart summarizing the overall analysis plan is displayed in Supplementary Fig. 1.

Study Populations

The Mount Sinai Bio*Me* Biobank is an EHR-linked biorepository comprising ~60,000 participants, all aged >18 years. Bio*Me* enrolls nonselectively from the Mount Sinai Health System, located in and serving the greater New York City area. During the enrollment process, participants complete a detailed demographic and lifestyle questionnaire, and they consent to link their de-identified EHR to their DNA and plasma information. Participants were excluded from analysis if their electronic medical records were not accessible because of privacy concerns. Genotyping was performed using the Illumina Global Screening Array or the Illumina Global Diversity Array. Imputation was performed using the National Heart, Lung and Blood Institute Trans-Omics for Precision Medicine (TOPMed) reference panel.

The Mass General Brigham (MGB) Biobank is another EHR-linked biobank based at the MGB hospital system in Boston, Massachusetts. There were ~40,000 participants with available genetic data at the time of this study. Genotyping was performed on DNA samples using the Illumina Multi-Ethnic Genotyping Array or the Infinium Global Screening Array. Imputation was performed using the TOPMed reference panel.

To optimize the type 1 diabetes polygenic scores (see below), we used a separate cohort from the UK Biobank (15). The UK Biobank is a large-scale prospective study with ~500,000 participants from the U.K. aged between 40 and 69 years. A majority of participants self-identified as White.

Race and Ethnicity

All categories of race and ethnicity were extracted from EHRs. Race and ethnicity values were ascertained using self-identification; however, we cannot exclude the possibility that race and ethnicity values in legacy versions of EHRs were assigned by other observers (such as clinic administrative staff members).

The two study sites had different options available for self-identification. Bio*Me* asked a single question ("What is your ancestry?"), whereas the MGB Biobank asked two separate questions ("What is your race?" and "What is your ethnicity?"). In order to harmonize the demographic information in the two biobanks, we applied standardized labels to each category, while acknowledging that these labels do not perfectly capture the information reported by each participant (Supplementary Table 1). Certain categories had very low numbers of participants and were combined together under the "Other" label.

Genetically Inferred Ancestry

As a sensitivity analysis, we also classified biobank participants by genetically inferred ancestry groups. We used principal component analysis to assign participants to one of six continental ancestry groups (African, American, Central/South Asian, East Asian, European, and Middle Eastern), following the method of the Pan-UK Biobank (16). We used a random forest classifier to determine the probability that a given individual matched a specific genetic ancestry group. Each individual was then assigned to the ancestry group that had the highest probability from the random forest classifier. If no ancestry group had a probability >50%, then the individual's genetic ancestry was left as "Unclassified."

Type 1 Diabetes Definitions

eMERGE Algorithm

We identified individuals in EHRs with type 1 diabetes using an electronic phenotyping algorithm developed at the Children's Hospital of Philadelphia for the eMERGE (Electronic Medical Records and Genomics) Consortium (17). The algorithm identifies individuals who have been assigned a type 1 diabetes–related ICD-9 or ICD-10 code and who have been prescribed insulin. It excludes individuals who have been prescribed type 2 diabetes medications or who have an ICD code for malignant cancer, cystic fibrosis, or drug-induced diabetes.

Medical Record Review

At each site, a trained medical reviewer performed manual record review for all individuals identified as having type 1 diabetes by the eMERGE algorithm. To confirm a diagnosis of type 1 diabetes, participants had to meet all of the following criteria, modified from (13):

- Diagnosis confirmed by an endocrinologist or primary care physician
- Current use of basal-bolus insulin or pump
- No secondary cause of diabetes listed in the medical record: gestational diabetes, checkpoint inhibitor use, glucocorticoid-induced diabetes, cystic fibrosis diagnosis, hemochromatosis, pancreatogenic diabetes, posttransplantation diabetes, maturity-onset diabetes of the young, or diagnosis of type 1.5 diabetes

Phenotypic Traits

BMI and hemoglobin A_{1c} (HbA_{1c}) values were extracted from EHR data. Median values were reported using the most

recent 5 years of data. For BMI, inpatient encounters were excluded because of wide variations in weight during inpatient admissions. Age and sex were self-reported. Age was defined at the time the data were accessed in the biobank, not at the time of diabetes diagnosis.

Statistics

To assess the accuracy of the type 1 diabetes classification algorithm, we calculated the positive predictive value (PPV), which is the proportion of individuals with putative type 1 diabetes confirmed to have true type 1 diabetes after manual medical record review. We used logistic regression models to assess the relationship between race and type 1 diabetes misdiagnosis, while simultaneously controlling for covariates such as age, sex, and BMI. We then meta-analyzed the results across the two biobanks using the R statistical package meta (18).

To evaluate the discriminatory power of the polygenic scores, we calculated the area under the curve (AUC) of the receiver operating characteristic curve, which evaluated type 1 diabetes status (case or control) using only the polygenic score as a predictor. Statistical comparisons between AUCs were performed using the DeLong test (19).

Polygenic Scores

We calculated two previously published polygenic scores to assess the genetic risk of type 1 diabetes. Both scores are restricted to significant polygenic scores (rsPSs), meaning that they include only a set of single-nucleotide polymorphisms (SNPs) reaching genome-wide significant association with type 1 diabetes (20). The first score (T1D-rsPS$_{EUR}$) was created in individuals with self-reported White or European ancestry (6,670 cases and 9,416 control participants) (13). The second score (T1D-rsPS$_{AFR}$) was created in individuals with self-reported Black or African ancestry (1,021 cases and 2,928 control participants) (14).

T1D-rsPS$_{EUR}$ included a weighted sum of 67 SNPs, where each risk allele was weighted by the log-odds of association from a genome-wide association study. The score also accounted for interactions between various HLA haplotypes by assigning different weights to distinct combinations of HLA alleles. Among the SNPs included in T1D-rsPS$_{EUR}$, certain variants were not available in the TOPMed imputation panel, so proxy SNPs were substituted in these instances (Supplementary Table 2). T1D-rsPS$_{AFR}$ included seven SNPs (five from chromosome 6 near HLA loci, one from chromosome 11, and one from chromosome 17), also weighted by the log-odds of association from a genome-wide association study.

Incorporation of Polygenic Scores in eMERGE Algorithm

To assess the impact of polygenic scores, we added an additional step to the eMERGE type 1 diabetes algorithm, where participants needed to have a polygenic score (T1D-rsPS$_{EUR}$ or T1D-rsPS$_{AFR}$) above a prespecified cutoff threshold to confirm the diagnosis of type 1 diabetes. The updated classification algorithms are denoted as eMERGE-rsPS$_{EUR}$ or eMERGE-rsPS$_{AFR}$.

To determine the optimal cutoff value for each polygenic score, we first implemented both scores in an independent population (UK Biobank). We identified the value for each polygenic score that maximized the Youden index (defined as j = sensitivity + specificity − 1). Because the UK Biobank had a very low number of non-White individuals with type 1 diabetes, we were not able to determine a cutoff for each self-reported racial group. Therefore, we did not restrict study participants by race, and we used the entire UK Biobank (which primarily comprises White participants) to determine the optimal cutoff value of both T1D-rsPS$_{EUR}$ and T1D-rsPS$_{AFR}$.

RESULTS

Implementation of eMERGE Type 1 Diabetes Algorithm

The two biobanks were similar in size and age distribution, with an average age of 58.7 years in Bio*Me* and 57.7 years in the MGB Biobank (Table 1). Bio*Me* had a higher proportion of Black and Hispanic participants, whereas MGB Biobank had a higher proportion of White participants.

The eMERGE type 1 diabetes algorithm identified 160 Bio*Me* participants and 172 MGB participants with putative type 1 diabetes (Table 1 and Supplementary Table 3). As expected, median HbA_{1c} was elevated among individuals with putative type 1 diabetes in both Bio*Me* (8.5% [69 mmol/mol]) and the MGB Biobank (8.1% [65 mmol/mol]); median HbA_{1c} was not available for the entire biobank because of missing values for a substantial proportion of participants.

Verification of Type 1 Diabetes Phenotype

To verify the type 1 diabetes phenotype, manual medical record reviews were conducted for all individuals with putative type 1 diabetes identified by the eMERGE algorithm. Each participant was then relabeled as having verified or misclassified type 1 diabetes. On manual record review, 122 of 160 participants with putative type 1 diabetes in Bio*Me* were confirmed to have type 1 diabetes (PPV 76%), as well as 116 of 172 participants in the MGB Biobank (PPV 67%).

To confirm that the manual record review process improved the classification of type 1 diabetes, we calculated polygenic scores for type 1 diabetes. When using the eMERGE type 1 diabetes algorithm to define case/control status, the AUC for T1D-rsPS$_{EUR}$ was 0.744 in the MGB Biobank, but the AUC improved to 0.875 after revising the type 1 diabetes case definition based on manual record reviews ($P = 2.5 \times 10^{-7}$) (Supplementary Fig. 2*A*). The results were similar in Bio*Me* (AUC 0.766 using the eMERGE type 1 diabetes algorithm to define case/control status and AUC 0.822 using manual record review) (Supplementary Fig. 2*B*), but the difference was not significant ($P = 0.059$). Findings were similar when restricting the analysis to self-reported White participants, which is the population in which T1D-rsPS$_{EUR}$ was developed (Supplementary Fig. 2*C* and *D*), and when using T1D-rsPS$_{AFR}$ (Supplementary Fig. 2*E* and *F*).

Analysis of Individuals With Misclassified Type 1 Diabetes

We next assessed whether participants with confirmed type 1 diabetes differed in clinical features from those who had been misclassified (Table 1 and Supplementary Table 4). In Bio*Me*, the average age of the individuals with confirmed type 1 diabetes was younger (47.8 years) compared with those with misclassified type 1 diabetes (65.0 years) ($P = 1.9 \times 10^{-9}$). A similar age difference was observed in the MGB Biobank (51.2 vs. 63.8 years; $P = 1.1 \times 10^{-6}$). Individuals with confirmed type 1 diabetes had a lower median BMI compared with those who had been misclassified (Bio*Me* 26.2 vs. 31.1 kg/m^2; $P = 3.1 \times 10^{-4}$; MGB Biobank 27.3 vs. 30.3 kg/m^2; $P = 5.1 \times$

Table 1—Baseline characteristics of MGB Biobank and Bio*Me* cohorts

	MGB Biobank			Bio*Me*		
	Entire biobank	Participants with putative type 1 diabetes (eMERGE algorithm)	Participants with verified type 1 diabetes (medical record review)	Entire biobank	Participants with putative type 1 diabetes (eMERGE algorithm)	Participants with verified type 1 diabetes (medical record review)
Total no. of participants	41,006	172	116	57,643	160	122
Self-identified race						
White	34,939	136	96	16,663	57	49
Black	2,101	20	11	11,443	29	25
Hispanic	1,270	4	1	19,524	50	35
Other*	1,511	6	4	10,013	24	13
Sex						
Female	22,418	87	63	33,389	92	72
Male	18,587	85	53	24,254	68	50
Age, years	57.7 ± 17.2	55.4 ± 16.4	51.2 ± 16.1	58.7 ± 17.9	51.3 ± 14.8	47.8 ± 13.8
BMI,† kg/m^2	28.6 ± 6.3	28.2 ± 5.8	27.3 ± 5.4	28.3 ± 6.6	27.2 ± 6.6	26.2 ± 5.3
HbA_{1c},† %	—	8.1 ± 1.6	8.0 ± 1.5	—	8.5 ± 1.9	8.5 ± 2.0
HbA_{1c},† mmol/mol	—	65 ± 17.5	64 ± 16.4	—	69 ± 20.8	69 ± 21.9

Data presented as *n* or mean ± SD. *Includes participants who selected any race other than the listed choices. †Median values over last 5 years.

10^{-3}). HbA_{1c} did not differ significantly between the two groups in either Bio*Me* or the MGB Biobank (Table 1).

The PPV of the eMERGE algorithm differed by self-reported race in both biobanks. In Bio*Me*, the eMERGE type 1 diabetes algorithm correctly identified 49 of the 57 White individuals with type 1 diabetes (86%), whereas only 73 of 103 non-White individuals (71%) were correctly classified (P = 0.03) (Table 2). Likewise, in the MGB Biobank, the eMERGE type 1 diabetes algorithm correctly identified 96 of 136 White individuals (71%), compared with only 16 of 30 non-White individuals (53%; P = 0.07). Across the two biobanks, these results remained significant after controlling for age, sex, and BMI. In a meta-analysis of Bio*Me* and the MGB Biobank, the odds of a non-White individual being misclassified as having type 1 diabetes was 3.45 (95% CI 1.54–7.69; $P = 2.6 \times 10^{-3}$), compared with a White individual (Fig. 1). Additionally, increased age and increased BMI were independently associated with higher odds of type 1 diabetes misclassification (Fig. 1).

Incorporation of Polygenic Scores

Next, we investigated whether type 1 diabetes polygenic scores could improve the identification of individuals with type 1 diabetes. We calculated two ancestry-specific polygenic scores that were restricted to genome-wide significant SNPs (T1D-rsPS$_{EUR}$ and T1D-rsPS$_{AFR}$), and we updated the eMERGE algorithm to include these scores. In Bio*Me*, for self-identified White individuals, inclusion of T1D-rsPS$_{EUR}$ improved the PPV from 86 to 100%, while inclusion of T1D-rsPS$_{AFR}$ improved the PPV to 97% (Fig. 2). For non-White individuals, the PPV improved from 71 to 93% with T1D-rsPS$_{EUR}$ and 86% with T1D-rsPS$_{AFR}$. The results were similar for the MGB Biobank; for instance, among self-identified White individuals, inclusion of T1D-rsPS$_{EUR}$ improved the PPV from 71 to 97%, whereas for non-White individuals, inclusion of T1D-rsPS$_{AFR}$ improved the PPV from 53 to 83% (Fig. 2).

However, while incorporating polygenic scores improved the PPV of the eMERGE type 1 diabetes algorithm, the sensitivity was reduced. For instance, there were a total of 96 White individuals with verified type 1 diabetes in the MGB Biobank but eMERGE-rsPS$_{EUR}$ identified only 65 individuals with verified type 1 diabetes.

We recognize that self-identified race is distinct from genetic ancestry, and the two labels cannot be used interchangeably; therefore, we also assessed the eMERGE type 1 diabetes algorithm after using principal component analysis to determine genetically inferred ancestry for MGB Biobank participants. Once again, we found that inclusion of T1D-rsPS$_{EUR}$ or T1D-rsPS$_{AFR}$ improved the PPV of the eMERGE type 1 diabetes algorithm (Supplementary Fig. 3).

Table 2—PPV of eMERGE type 1 diabetes algorithm across racial groups

	PPV* of eMERGE algorithm, %	
Self-identified race	MGB Biobank	Bio*Me*
White	70.6	86.0
Black	55.0	86.2
Hispanic	25.0	70.0
Other†	66.7	54.2
All non-White groups	53.3	70.9
Total	67.4	76.3

*Proportion of participants with putative type 1 diabetes whose phenotype was verified after manual medical record review. †Includes participants who selected any race other than the listed choices.

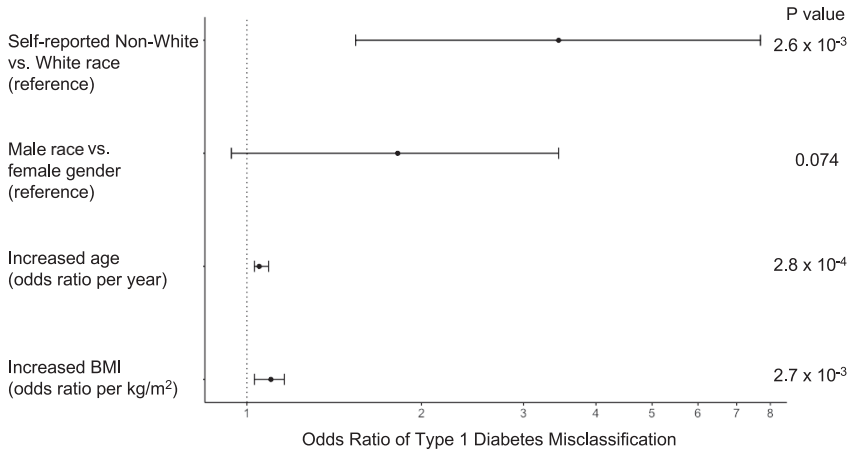


Figure 1—Meta-analysis of eMERGE type 1 diabetes classification algorithm performance compared with manual medical record review. The forest plot demonstrates how different factors affect the likelihood of misclassification by the eMERGE type 1 diabetes algorithm in a meta-analysis of participants from the MGB Biobank and Bio*Me*. Odds ratios were obtained from a single logistic regression model that simultaneously controlled for race, sex, age, and BMI. The 95% CI is displayed for each data point.

CONCLUSIONS

We analyzed an automated algorithm developed by the eMERGE Consortium to identify individuals with type 1 diabetes in large biobanks. Using manual medical record reviews as a gold standard, we found that the eMERGE algorithm was moderately effective for identifying individuals with type 1 diabetes in a racially diverse adult cohort, with a combined PPV of 72% across Bio*Me* and the MGB Biobank. However, there was a clear bias in the performance of the eMERGE type 1 diabetes algorithm across race, with worse performance in non-White individuals.

There are various possible explanations for the disparity in performance between racial groups. The eMERGE algorithm incorporates diagnosis codes, which are entered by clinicians during routine clinical care and may be subject to bias. Furthermore, the eMERGE algorithm was developed in a pediatric population, with White children comprising the majority (H. Qu, personal communication), so the algorithm may not be optimized for racially diverse adult cohorts. For instance, the eMERGE algorithm excludes individuals who have been treated with type 2 diabetes medications, but the use of type 2 diabetes medications differs between children and adults (21) as well as between racial groups (22,23).

In addition, the heterogeneity of diabetes across different populations may contribute to the lower performance of the eMERGE type 1 diabetes algorithm in non-White individuals. On average, compared with White individuals, Black and Hispanic individuals have an earlier onset of type 2 diabetes (24); consequently, young adults in these populations may be misdiagnosed as having type 1 diabetes. Furthermore, although individuals who develop diabetic ketoacidosis are commonly diagnosed with type 1 diabetes, these individuals may also have ketosis-prone diabetes, an atypical form of diabetes that also involves ketoacidosis but is distinct from type 1 diabetes. Ketosis-prone diabetes was initially described in individuals with African ancestry, but it has also been described in Hispanic, Asian, and other populations (25,26). Therefore, the decreased performance of the eMERGE type 1 diabetes algorithm in non-White individuals may be related to the presence of ketosis-prone diabetes or another form of atypical diabetes in these populations, although ketosis-prone diabetes is poorly understood and remains an active area of investigation. Overall, further work should test the performance of the eMERGE algorithm as well as additional automated type 1 diabetes classification algorithms in other racially diverse cohorts.

We demonstrated that adding type 1 diabetes polygenic scores to the eMERGE type 1 diabetes algorithm (eMERGE-rsPS$_{EUR}$ and eMERGE-rsPS$_{AFR}$) can help identify individuals with true type 1 diabetes and reduce the disparity in misclassification rates among self-reported racial groups. As increasing numbers of biobanks incorporate genetic information, this strategy can be used to identify individuals with type 1 diabetes in biobanks for additional research studies. Of note, although autoantibodies such as GAD65 can be used to confirm the

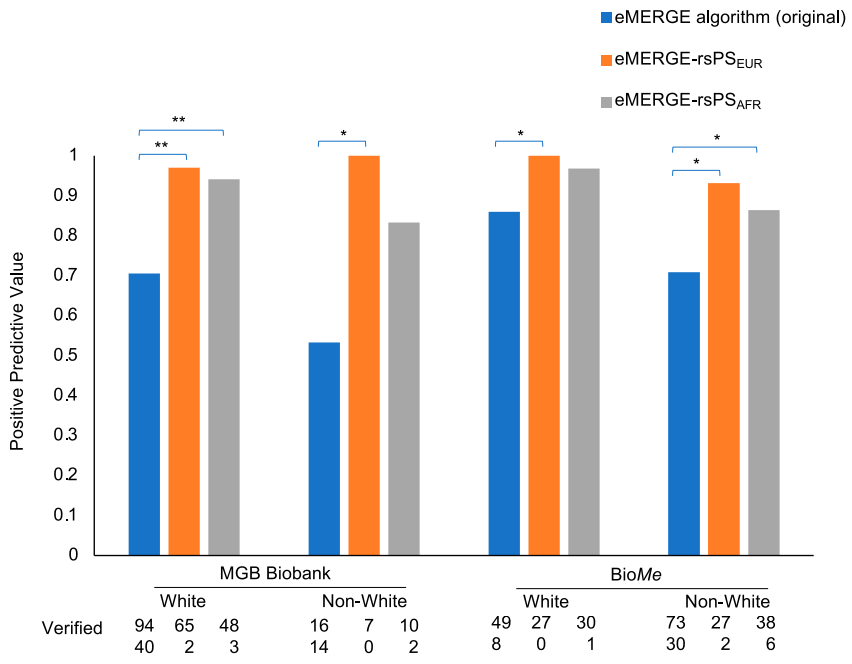


Figure 2—Improvement of eMERGE type 1 diabetes algorithm with inclusion of polygenic scores. The PPV of specified type 1 diabetes algorithms is shown for individuals in the MGB Biobank or Bio*Me*, as classified by self-reported race (White vs. non-White). Values are displayed for the original eMERGE type 1 diabetes algorithm, as well as for modified versions that also require individuals to have a polygenic score greater than a specified cutoff value (T1D-rsPS$_{EUR}$ (13) or T1D-rsPS$_{AFR}$ (14)). The optimal cutoff value was identified in a separate cohort (UK Biobank). The raw number of individuals with verified and misclassified type 1 diabetes is displayed beneath the graph. Statistical significance was assessed with two-sample test of proportions. *$P < 0.05$, **$P < 0.005$.

diagnosis of type 1 diabetes, in practice the number of biobank participants with autoantibody testing is very low. Therefore, genetic predisposition for type 1 diabetes (as captured by the polygenic score) offers a useful alternative with which to identify individuals with type 1 diabetes.

Previous studies have shown that type 1 diabetes polygenic scores developed in a single population can be applied to a more diverse population, but the predictive power is variable across race and ethnicity (27–30). Notably, self-identified race is not interchangeable with genetic ancestry; however, because race and ethnicity are correlated with genetic ancestry (31), ancestry-specific polygenic scores may perform differently in different racial groups. We found that among self-identified White participants, T1D-rsPS$_{EUR}$ showed greater discriminatory power compared with T1D-rsPS$_{AFR}$. For non-White participants, eMERGE-rsPS$_{EUR}$ had the highest PPV, but this was at the expense of reduced sensitivity. For instance, in Bio*Me*, only 27 non-White individuals with verified type 1 diabetes were identified using eMERGE-rsPS$_{EUR}$ compared with 38 individuals when using eMERGE-rsPS$_{AFR}$. Therefore, choosing the optimal polygenic score requires a tradeoff between optimizing sensitivity versus maximizing PPV.

One important limitation of this study is that we focused on the PPV of the eMERGE type 1 diabetes algorithm, but we did not assess the negative predictive value because of the limited feasibility of performing manual medical record reviews for thousands of individuals. Because the prevalence of type 1 diabetes is highest among White individuals, it is possible that type 1 diabetes is underdiagnosed in other populations. This has significant implications for public health because failure to recognize type 1 diabetes can lead to worse glycemic control and increased rates of diabetic ketoacidosis.

Another limitation to note is the small sample size included in this study. Type 1 diabetes accounts for only 5–10% of all diabetes cases, and the additional exclusion of type 2 diabetes medications further decreased the available participants. This affected the sample size of non-White participants in the MGB Biobank, where there were $<$5,000 non-White participants and a very limited number of individuals with type 1 diabetes. For instance, among Hispanic participants in the MGB Biobank, the PPV of the eMERGE type 1 diabetes algorithm was notably low at 25%, but this corresponded to just one of four individuals in this subgroup (Table 2).

Additionally, although all non-White biobank participants were analyzed together to maximize sample size, they represent multiple populations with diverse ancestry. Very few type 1 diabetes polygenic scores have been developed in non-White populations (32). T1D-rsPS$_{AFR}$ was developed for individuals with self-reported Black or African ancestry; however, additional studies are needed to develop type 1 diabetes polygenic scores in other populations, such as Hispanic individuals. Recent work has shown that modification of T1D-rsPS$_{EUR}$ with the addition of four African-specific variants can improve the predictive power in individuals with African ancestry (28). Ongoing efforts are underway to develop multiancestry type 1 diabetes polygenic scores (33) using meta-analyses that incorporate participants from multiple populations (34). Future efforts may classify individuals according to genetically inferred ancestry groups; notably, however, the disparity in the eMERGE type 1 diabetes algorithm was present when classifying individuals by self-reported race, irrespective of genetic ancestry.

In this study, both participating biobanks are hospital based and are subject to selection biases, such as Berkson bias. This bias arises when a sample is taken from a subpopulation and not the overall general population. To be included in the current study, participants were required to have some affiliation with either the Mount Sinai or MGB health systems, biasing the study to be less healthy than the general public. Furthermore, within the hospital cohorts, it is possible that certain populations are more likely to provide consent to use genomic data. This makes these results less generalizable to the general public.

Overall, we demonstrated an important disparity in the performance of an automated classification algorithm to detect individuals with type 1 diabetes, and we identified a potential solution by incorporating polygenic scores. Further work is needed to elucidate the sources of this disparity. Accurate diagnosis of diabetes subtypes in non-White populations is likely to be a critical component for reducing disparities in diabetes outcomes. Future multiancestry type 1 diabetes polygenic scores may help to reduce this disparity even further.

Acknowledgments. The authors thank the D-PRISM (Diabetes Polygenic Risk Scores in Multiple Ancestries) study site of the PRIMED (Polygenic Risk Methods in Diverse Populations) Consortium for helpful comments on this project.

Funding. A.J.D. was supported by National Institutes of Health (NIH)/National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) grant T32DK007028. L.S. was supported by NIH/NIDDK grant F30DK130576. J.M.M., A.K.M., and M.S.U. were supported by NIH/National Human Genome Research Institute grant U01HG011723. J.C.F. was supported by NIH/National Heart, Lung, and Blood Institute grant K24 HL157960. M.S.U. was supported by NIH/NIDDK grant K23DK114551 and the Massachusetts General Hospital Transformative Scholar Award.

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions. A.J.D., L.S., J.C.F., and M.S.U. designed the study. A.J.D. and L.S. performed the analysis and wrote the initial draft of the manuscript. T.D.M. and J.M.M. assisted with implementation and analysis of polygenic scores in the MGB Biobank. A.K.M. provided input on the study design and analysis plan. J.C.F., R.J.F.L., and M.S.U. supervised the study and edited the manuscript. All authors approved the final version of the manuscript. M.S.U. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. Parts of this study were presented in abstract form at the 82nd Scientific Sessions of the American Diabetes Association, New Orleans, LA, 3–7 June 2022.

References

1. Denny JC, Rutter JL, Goldstein DB, et al.; All of Us Research Program Investigators. The "All of Us" research program. N Engl J Med 2019;381:668–676
2. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005;40:1620–1639
3. Pendergrass SA, Crawford DC. Using electronic health records to generate phenotypes for research. Curr Protoc Hum Genet 2019;100:e80
4. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016;23(e1):e20–e27
5. Lethebe BC, Williamson T, Garies S, et al. Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study. CMAJ Open 2019;7:E246–E251
6. Klompas M, Eggleston E, McVetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. Diabetes Care 2013;36:914–921
7. Schroeder EB, Donahoo WT, Goodrich GK, Raebel MA. Validation of an algorithm for identifying type 1 diabetes in adults based on electronic health

record data. Pharmacoepidemiol Drug Saf 2018;27:1053–1059
8. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. Clin Epidemiol 2016;8:373–380
9. Lo-Ciganic W, Zgibor JC, Ruppert K, Arena VC, Stone RA. Identifying type 1 and type 2 diabetic cases using administrative data: a tree-structured model. J Diabetes Sci Technol 2011;5:486–493
10. Weisman A, Tu K, Young J, et al. Validation of a type 1 diabetes algorithm using electronic medical records and administrative healthcare data to study the population incidence and prevalence of type 1 diabetes in Ontario, Canada. BMJ Open Diabetes Res Care 2020;8:e001224
11. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19:212–218
12. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178:1544–1547
13. Sharp SA, Rich SS, Wood AR, et al. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. Diabetes Care 2019;42:200–207
14. Onengut-Gumuscu S, Chen WM, Robertson CC, et al.; SEARCH for Diabetes in Youth; Type 1 Diabetes Genetics Consortium. Type 1 diabetes risk in African-ancestry participants and utility of an ancestry-specific genetic risk score. Diabetes Care 2019;42:406–415
15. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–209
16. Pan-UK Biobank. Pan-ancestry genetic analysis of the UK Biobank. Accessed 25 August 2022. Available from https://pan.ukbb.broadinstitute.org
17. Qu H, Roizen J, Mentch F, et al. CHOP. Type 1 diabetes. Accessed 26 May 2022. Available from https://phekb.org/phenotype/1548
18. Balduzzi S, Rucker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health 2019;22:153–160
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845
20. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocr Rev 2019;40:1500–1520
21. American Diabetes Association Professional Practice Committee. 14. Children and adolescents: standards of medical care in diabetes-2022. Diabetes Care 2022;45(Suppl. 1):S208–S231
22. Elhussein A, Anderson A, Bancks MP, et al.; Look AHEAD Research Group. Racial/ethnic and socioeconomic disparities in the use of newer diabetes medications in the Look AHEAD study. Lancet Reg Health Am 2022;6:100111
23. Eberly LA, Yang L, Essien UR, et al. Racial, ethnic, and socioeconomic inequities in glucagon-like peptide-1 receptor agonist use among patients with diabetes in the US. JAMA Health Forum 2021;2:e214182
24. Wang MC, Shah NS, Carnethon MR, O'Brien MJ, Khan SS. Age at diagnosis of diabetes by race and ethnicity in the United States from 2011 to 2018. JAMA Intern Med 2021;181:1537–1539
25. Balasubramanyam A, Nalini R, Hampe CS, Maldonado M. Syndromes of ketosis-prone diabetes mellitus. Endocr Rev 2008;29:292–302
26. Lebovitz HE, Banerji MA. Ketosis-prone diabetes (Flatbush diabetes): an emerging worldwide clinically important entity. Curr Diab Rep 2018;18:120
27. Perry DJ, Wasserfall CH, Oram RA, et al. Application of a genetic risk score to racially diverse type 1 diabetes populations demonstrates the need for diversity in risk-modeling. Sci Rep 2018;8:4529
28. Qu HQ, Qu J, Glessner J, et al. Improved genetic risk scoring algorithm for type 1 diabetes prediction. Pediatr Diabetes 2022;23:320–323
29. Oram RA, Sharp SA, Pihoker C, et al. Utility of diabetes type-specific genetic risk scores for the classification of diabetes type among multiethnic youth. Diabetes Care 2022;45:1124–1131
30. Kaddis JS, Perry DJ, Vu AN, et al. Improving the prediction of type 1 diabetes across ancestries. Diabetes Care 2022;45:e48–e50
31. Banda Y, Kvale MN, Hoffmann TJ, et al. Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. Genetics 2015;200:1285–1295
32. Redondo MJ, Gignoux CR, Dabelea D, et al. Type 1 diabetes in diverse ancestries and the use of genetic risk scores. Lancet Diabetes Endocrinol 2022;10:597–608
33. Mercader JM, Ng MCY, Manning AK, Rich SS. Predicting diabetes risk in diverse populations: what next? Lancet Diabetes Endocrinol 2021;9:808–810
34. Robertson CC, Inshaw JRJ, Onengut-Gumuscu S, et al.; Type 1 Diabetes Genetics Consortium. Fine-mapping, trans-ancestral and genomic analyses identify causal variants, cells, genes and drug targets for type 1 diabetes. Nat Genet 2021;53:962–971