Journal of the American Medical Informatics Association, 2023, **30(12)**, 1915–1924
https://doi.org/10.1093/jamia/ocad154
Advance access publication 3 August 2023
Research and Applications

Research and Applications

The value of parental medical records for the prediction of diabetes and cardiovascular disease: a novel method for generating and incorporating family histories

Yuval Barak-Corren[1]*, David Tsurel[1,2,3], Daphna Keidar[1,2], Ilan Gofer[2], Dafna Shahaf[3], Maya Leventer-Roberts[2,4,5], Noam Barda[2], and Ben Y. Reis[1,6]

[1]Predictive Medicine Group, Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, USA, [2]Clalit Research Institute, Ramat Gan, Israel, [3]The Hebrew University of Jerusalem, Jerusalem, Israel, [4]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, New York, USA, [5]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, New York, USA, and [6]Harvard Medical School, Boston, Massachusetts, USA

***Corresponding Author:** Yuval Barak-Corren, MD, MSc, Predictive Medicine Group, Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA 02215, USA; yuval.barakcorren@childrens.harvard.edu

ABSTRACT

Objective: To determine whether data-driven family histories (DDFH) derived from linked EHRs of patients and their parents can improve prediction of patients' 10-year risk of diabetes and atherosclerotic cardiovascular disease (ASCVD).

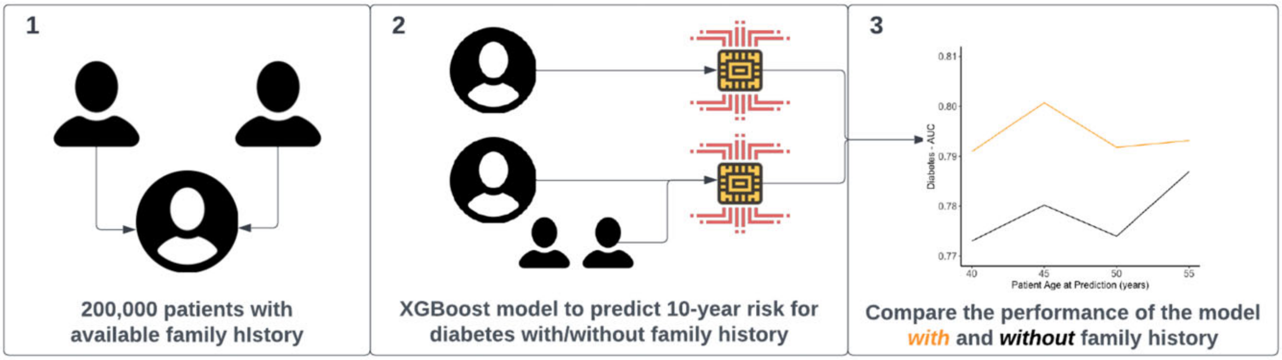
Materials and Methods: A retrospective cohort study using data from Israel's largest healthcare organization. A random sample of 200 000 subjects aged 40–60 years on the index date (January 1, 2010) was included. Subjects with insufficient history (<1 year) or insufficient follow-up (<10 years) were excluded. Two separate XGBoost models were developed—1 for diabetes and 1 for ASCVD—to predict the 10-year risk for each outcome based on data available prior to the index date of January 1, 2010.

Results: Overall, the study included 110 734 subject-father-mother triplets. There were 22 153 cases of diabetes (20%) and 11 715 cases of ASCVD (10.6%). The addition of parental information significantly improved prediction of diabetes risk ($P < .001$), but not ASCVD risk. For both outcomes, maternal medical history was more predictive than paternal medical history. A binary variable summarizing parental disease state delivered similar predictive results to the full parental EHR.

Discussion: The increasing availability of EHRs for multiple family generations makes DDFH possible and can assist in delivering more personalized and precise medicine to patients. Consent frameworks must be established to enable sharing of information across generations, and the results suggest that sharing the full records may not be necessary.

Conclusion: DDFH can address limitations of patient self-reported family history, and it improves clinical predictions for some conditions, but not for all, and particularly among younger adults.

GRAPHICAL ABSTRACT


1
200,000 patients with available family hIstory
2
XGBoost model to predict 10-year risk for diabetes with/without family history
3
0.81
0.80
0.79
0.78
0.77
Diabetes - AUC
40
45
50
55
Patient Age at Prediction (years)
Compare the performance of the model *with* and *without* family history

Overall, 110,734 Father-Mother-Child triplets with linked medical records met the study inclusion criteria.

A machine-learning algorithm was trained to predict the 10-year risk for developing new-onset diabetes and cardiovascular disease, with and without the data from the parents' linked medical records.

Family history (FH) significantly improved the prediction of diabetes, but not of cardiovascular disease. FH was more significant among the younger patients (<40).

Received: 1 May 2023. **Revised:** 16 July 2023. **Editorial Decision:** 21 July 2023. **Accepted:** 24 July 2023

Key words: family history, clinical prediction, cardiovascular disease, diabetes, health informatics

INTRODUCTION

Family medical history, often abbreviated as Family History (FH), is an essential predictor of disease risk,[1,2] encapsulating information on a patient's genetic susceptibility to disease, and often also on a patient's social and physical environment, as well as health-related behaviors.[3] Positive FH (ie, a record of a disease having occurred in a family) is a well-established risk factor for both rare and common conditions that, in aggregate, account for the majority of healthcare costs, morbidity, and mortality in developed countries.[4–8] A positive FH can also increase a patient's perceived risk for disease, a critical factor for changing health behaviors.[9] Indeed, studies have shown that knowledge of FH has led to more frequent health screenings.[10]

Despite its importance, FH is not widely available in today's clinical settings.[11,12] Data from the Electronic Medical Records and Genomics (eMERGE) Network[13] showed that only 2 of 5 major healthcare systems captured any FH information whatsoever in their electronic medical records, and that over 70% of the records were incomplete or missing in the centers that did collect this information.[14] FH is difficult to collect from patients due to the time limitations of the clinical encounter[1,4] and the patient's inherently incomplete knowledge of their own FH[1,4,15]: Patients often do not have complete or reliable knowledge about their immediate (first degree) family members' health history, and typically know even less about more distant relatives.[16] Studies have also shown that patient-reported family histories are typically recorded only after the patient has already been diagnosed with a particular condition, diminishing their value for advanced disease risk prediction.[17] The accuracy of patient-reported FH has also been found to depend on the patient's own health status.[18]

Recent efforts focused on streamlining the process of collecting family histories from patients, including the Surgeon General's *My Family Health Portrait*[19] and many others,[4,20,21] have made important progress, but continue to face significant challenges.[20,22] These systems remain limited by time constraints and patients' incomplete knowledge of their own FH.[1,23] Many of them also face difficulties transferring the information they collect to the EHR.[20]

Even when FH is available, many health professionals do not incorporate this information into their clinical decisions.[1,2,4,12,23] This is partly due to the fact that it is difficult for physicians to interpret complex family histories and calculate clinical risks during the short clinical encounter.[1] Computerized risk assessment of family histories can help overcome these challenges by integrating many types of data when performing risk calculations.[24] The supplementary FH information might enable physicians to identify high-risk patients, undetected by patient-specific factors. This could facilitate more personalized doctor-patient discussions and enhance patient motivation for health behavior changes.[9,10] Therefore, FH presents a largely untapped opportunity for health information technology innovation and intervention.[4]

To address these challenges and to overcome the inherent limitations of patient-reported FH, we sought to examine a data-driven approach for capturing FH. We analyzed EHR data from a large healthcare system database containing linked family groups of parents and offspring.[25] We focused on predicting risks of developing atherosclerotic cardiovascular disease (ASCVD) and diabetes (type 2 diabetes mellitus), 2 leading causes of morbidity and mortality for which early screening and lifestyle modifications have the potential to improve prognosis and even prevent the development of disease altogether.[26] We compared the performance of models that predict new-onset ASCVD and diabetes based only on the subject's EHR versus models that also take into account data from the EHRs of the subject's parents. To help guide potential future interventions, we sought to identify specific groups for whom FH might be particularly beneficial or less useful, examining whether the predictive value of FH varied by the age or sex of the subject, or age or sex of the parent. We also compared the specific features in the father's and mother's health records which had the highest predictive value.

MATERIALS AND METHODS

Clalit Health Services (CHS) is the largest of 4 payer-provider health care delivery systems in Israel, with CHS serving over 4.7 million members—over half of the Israeli population. The CHS database includes more than 2 million nuclear family groups in which both parents and their offspring are members of CHS. Annual membership attrition is below 2% so there is minimal loss to longitudinal follow up.[27] CHS members are registered using their national identification number which is also linked to their birth records through the Ministry of Health and residential information through the Ministry of Interior. The birth records provide a link between the patient and his/her parents' national identification numbers, thus enabling a link between the patient's and parents' CHS medical records. As such, the CHS data warehouse stores comprehensive and validated clinical and demographic information on its entire patient population.

The data were deidentified prior to analysis, removing identifiable information and randomly offsetting the dates. We extracted a randomly selected sample of 200 722 subjects from the CHS data warehouse according to the following inclusion criteria: (1) subject was aged 40–60 years on the index date of January 1, 2010, and (2) both of the subject's parents were members of CHS for at least 1 year from 2003 onward. The age group of 40–60 years was selected since it had a significant concentration of both ASCVD and diabetes incidence according to a previous study conducted using the same dataset.[25] Subjects who met the case-definition prior to the index date, who were covered by the health system for less than 1 year prior to the index date, or who had insufficient follow-up time to ascertain the case definition (<10 years follow-up) were excluded. Within the selected sample of 200 722, if 2 or more siblings from the same family were found in the sample, only the oldest of these siblings was included as subject in the study. For each *subject-mother-father* triplet, we extracted the complete medical record, including demographics, diagnoses, laboratory tests, medications, procedures, and hospitalizations for all 3 persons in the triplet. Of the laboratory tests, 2 tests were converted into categorical variables: glucose (<5.56, 5.56–6.94, 6.95–11.11, and >11.11 mmol/L or <100, 100–125, 126–200, and >200 mg/dL) and hemoglobin A1C (<39, 39–47, and ≥48 mmol/mol or <5.7%, 5.7%–6.4%, and ≥6.5%). Missing variables were recorded as "not available" and included in the analysis.

Ethics

The study was approved by the CHS Institutional Review Board. Reporting followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist.[28]

Outcomes and case definitions

We examined 2 outcomes for the study subjects: New onset diabetes within 10 years of the index date, and new onset ASCVD within 10 years of the index date.

ASCVD case definition

An ASCVD event was defined as either acute myocardial infarction, stroke, or fatal coronary artery disease (CAD) in a patient without any history of cardiovascular disease (CVD). A fatal CAD was defined as death within 1 year of the CAD diagnosis, as recorded in the patient's EHR. Following the design of Ward et al,[29] a broader case definition than ASCVD was used for the exclusion of patients with pre-existing CVD and for the definition of positive FH of CVD. This case definition encompassed both ASCVD and any of the following: diagnosis of atrial fibrillation, heart failure, or other CVD (Supplementary Appendix S1).

Diabetes case definition

Diabetes was defined as a diagnosis of diabetes accompanied by laboratory evidence for diabetes. Patients who met only one of these criteria (only diagnosis or only labs) were excluded from the study. The same case definition was used both for subjects and for their parents (to define positive FH of diabetes). The laboratory criteria were based on the American Diabetes Association's recommendations,[30] requiring at least one of the following 3 conditions:

1) Hemoglobin A1C levels of 48 mmol/mol (6.5%) and over.
2) Fasting glucose levels of over 125 mg/dL (information on fasting was not readily available, thus we only included tests taken before 9 AM).
3) Glucose levels of over 11.11 mmol/L (200 mg/dL) 2 h after an Oral Glucose Tolerance Test (OGTT) of 75 g glucose.

Lab results were obtained both from inpatient and outpatient encounters and a single abnormal lab result was sufficient to meet the laboratory criteria.

Diagnostic codes used for ASCVD and diabetes case definitions

Diagnostic criteria for both ASCVD and diabetes were based on ICD-9 codes validated in prior studies,[29,31,32] and on free-text search within uncoded diagnoses (eg, a search for the term "atherosclerosis" within the textual description of the diagnoses). ICD-10 codes were not in use at CHS during the study period and were thus not included in the case definition. Free-text search was only conducted within the diagnosis field and not within other fields such as the clinician's notes. A detailed description of the specific codes and search terms used is presented in Supplementary Appendix S1.

Preliminary analysis

A preliminary analysis of all demographic and clinical features was performed, both for the entire cohort as well as for the 2 subcohorts of ASCVD and diabetes case-subjects. Comparisons between cohorts were conducted using the Mann-Whitney-Wilcoxon test for continuous variables and chi-square for categorical variables. Risk ratios (RRs) were calculated to compare binary variables. To study the relation between FH and offspring's disease, adjusted risk ratios (ARRs) were calculated using a Poisson regression. These ARRs were adjusted by the following patient related confounders: other parent FH, age, sex, marital status, area of residence (periphery rank, urban/rural, geographic administrative code), body mass index (BMI), and presence of hypertension (yes/no). These potential confounders (and all other available variables in our data) were also addressed in the main analysis of this study—the multivariate predictive model below.

Predictive modeling

Two separate models were developed—1 for ASCVD and 1 for diabetes—to predict the 10-year risk for each outcome based on data available prior to the index date of January 1, 2010. The models were developed using XGBoost, an open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that predicts a target variable by combining the estimates of a set of simpler models (or trees). XGBoost has been shown to provide superior performance compared to other models in similar studies.[33–35] Moreover, it is computationally efficient, and thus well suited for the current prediction task which incorporates the medical records of 3 subjects for each prediction. For each outcome, the cohort was randomly divided into 2 subsets—an 80% training set and a 20% testing set. Hyperparameter tuning was performed on the training set using 20 iterations of a random-search within a predefined parameter space and with 5-fold cross-validation (within the training set). Using these hyperparameters, separate analyses were performed for subjects who were 40–44, 45–49, 50–54, 55–60 years old on the index date. For each age group, a separate model was derived from the training set and validated on the testing set. A more detailed description of XGBoost and the modeling process can be found in the Supplementary Material (Supplementary Appendix S2). The Supplementary Material also includes a literature review of known confounders for these prediction tasks, alongside a summary of which of these confounders was available in our data.

For each condition (ASCVD and diabetes) and for each age group, we compared 6 different models: (1) a prediction based solely on the subject's own EHR data; (2–4) predictions based only on the father's, mother's or both parents' EHR data, respectively; (5) a prediction based on the subject's data along with 2 binary variables indicating whether the father or mother met the case definition, and (6) a prediction based on the subject's data and both parents' full EHR data.

For each model, using the predictions generated for the testing set, we calculated the following performance metrics: Positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Using the *pROC* package in R,[36] 95% confidence intervals for AUC values were calculated using DeLong's method, and *P* values were derived for comparing AUCs of different models. For the other performance metrics (eg, sensitivity, PPV, etc.), bootstrap resampling with 1000 repetitions was used.[37] For each age group, and for each outcome, we compared the performance of the different model variations.

RESULTS

The records of 200 722 subjects were extracted from the CHS electronic medical records. Of these, 89 988 (45%) nonoldest siblings were excluded to keep only the oldest sibling in each family that was selected in the sample. Thus, a total of 110 734 subjects were included (Figure 1). The included cohort of subjects was composed of 54 161 (49%) men and 56 573 (51%) women, with a median age of 48.1 years on the index date (IQR 43.7–52.6, Table 1). The median duration of longitudinal EHR coverage available for each subject (from first to last documentation in the EHR) was 22.6 years (IQR 21.6–34.8). Overall, 3 104 109 person years were available for the subjects. The EHRs for the parents of each subject were also extracted. The median age of the parents was 75 years on the index date (IQR 70–80). Parent data coverage included a median of 22.7 years per parent, adding to a total of 9 717 976 person years for the entire study population.

Diabetes—preliminary analysis

The subjects included 22 153 cases of diabetes (20%). Median age at diagnosis, as recorded in the EHR, was 50.3 (IQR 45.7–54.0). 62.3% of diabetic subjects were male (vs 48.9% males in overall study population, Table 1). The number of subjects eligible for inclusion in the study, and the number of positive cases, both varied between the different age groups (Supplementary Appendix S3): As age increased, the number of eligible subjects decreased, and the percent of positive cases increased. Among individuals who were 40–45 years old on the index date, 27 962 subjects were eligible for prediction, of which 2702 (9.7%) were diagnosed with diabetes during the 10-year follow-up. Within the 45–50, 50–55, and 55–60 years age groups, there were 3138 (13.4%), 3542 (16.5%), and 1629 (18%) subjects, respectively, who met the diabetes case definition during the follow-up period (Supplementary Table S1 in Supplementary Appendix S3). Of those who developed diabetes, the median time from prediction to the index-event was between 4.2 and 5.7 years, depending on the age group.

The ARR for diabetes in our cohort was 1.51 (95% CI 1.29–1.78) if the father had diabetes and 1.75 (95% CI 1.49–2.06) if the mother had diabetes (Table 2). A parental history of disease was more predictive of a subject's risk when subjects or their parents were diagnosed at a younger age.

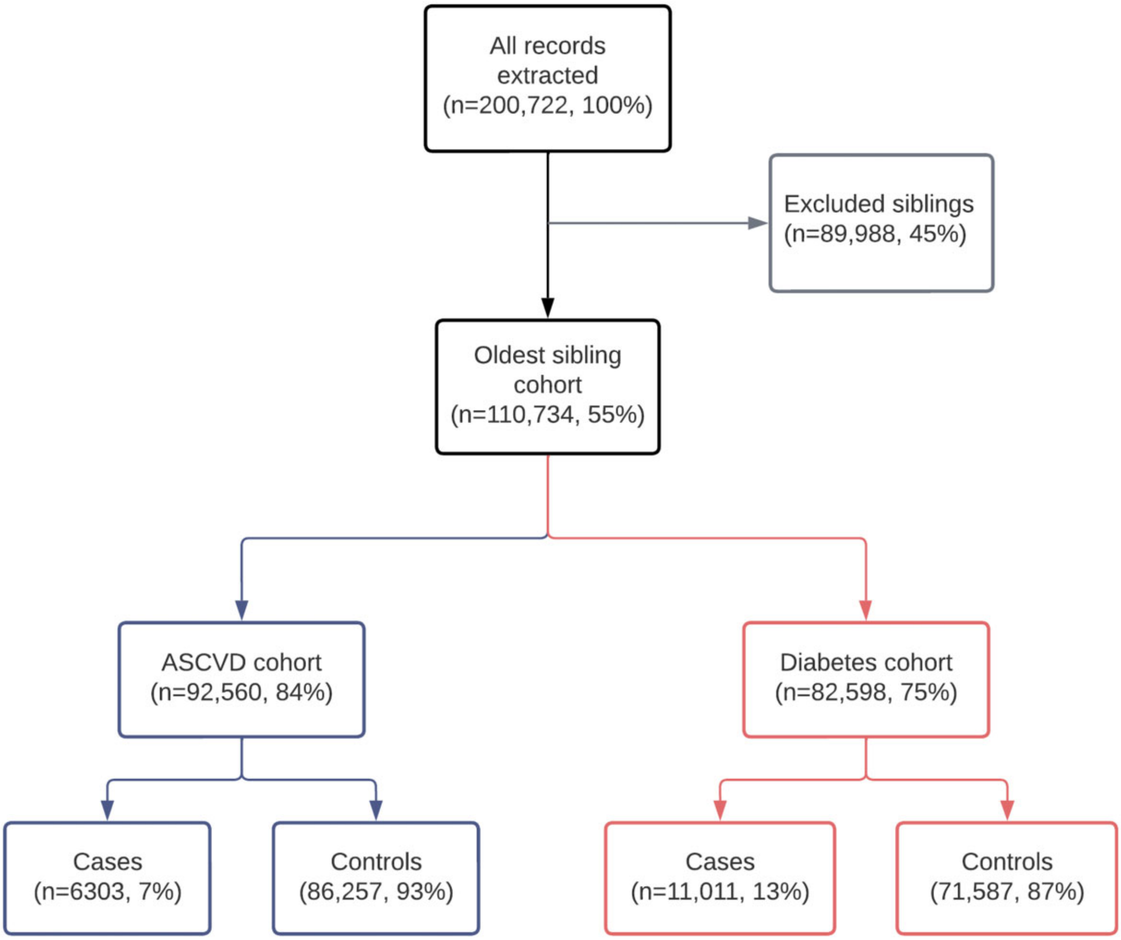

All records extracted (n=200,722, 100%)
Excluded siblings (n=89,988, 45%)
Oldest sibling cohort (n=110,734, 55%)
ASCVD cohort (n=92,560, 84%)
Diabetes cohort (n=82,598, 75%)
Cases (n=6303, 7%)
Controls (86,257, 93%)
Cases (n=11,011, 13%)
Controls (71,587, 87%)

Figure 1. Study flow chart. For each outcome, only patients that did not meet the case-definition prior to the prediction date (January 1, 2010) and with >10-year follow-up were eligible for prediction.

Table 1. Demographics of all study cohort, ASCVD cases, and diabetes cases

	All Patients 110 734 (100%)	ASCVD 11 715 (10.6%)	Diabetes 22 153 (20%)
Age (years)	48.1 [43.7–52.6]	51.0 [46.5–54.5]	50.3 [45.7–54]
Ethnicity			
Ashkenazi Jew	21 541 (19.5%)	2437 (20.8%)	4489 (18.2%)
Arab[a]	15 058 (13.6%)	1948 (16.6%)	4611 (18.7%)
Sephardi Jew	40 933 (37.0%)	4607 (39.3%)	7865 (32.0%)
Unknown/other	33 202 (30.0%)	2723 (23.2%)	5188 (23.4%)
Sex (% female)	56 573 (51.1%)	3258 (27.8%)	9274 (37.7%)
Marital status (married)	72 895 (65.8%)	7686 (65.6%)	14 807 (60.2%)
Number of children	3 [3–3]	3 [2–4]	3 [2–4]
Periphery rank	7 [6–9]	7 [5–9]	7 [5–9]
Urban dwelling	91 255 (82.4%)	9267 (79.1%)	18 366 (74.6%)
Age at diagnosis		52.8 [47.4–58.3]	50.4 [44.8–55.4]
Deceased[b]	3469 (3.1%)	1062 (9.1%)	1288 (5.8%)

[a] The "Arab" ethnicity-category describes all Israeli patients of Arab descent; no information was available to discern between the different subpopulations within this category.

[b] Record of mortality anytime within the extracted medical history. Continuous numbers are shown as median and IQR.

Table 2. Risk ratio with 95% confidence intervals for a subject meeting the case definition if one, either, or both parents met the case definition

	Diabetes RR			ASCVD RR		
Met case definition	**All subjects** **(*n* = 110 734)**	**Male subjects** **(*n* = 54 161)**	**Female subjects** **(*n* = 56 573)**	**All subjects** **(*n* = 110 734)**	**Male subjects** **(*n* = 54 161)**	**Female subjects** **(*n* = 56 573)**
Father	1.65[a]	1.64	1.67	1.48[a]	1.45	1.60
	[1.61–1.69]	[1.59–1.69]	[1.61–1.73]	[1.42–1.55]	[1.37–1.52]	[1.46–1.75]
Mother	1.97[a]	1.84	2.17	1.57[a]	1.55	1.62
	[1.92–2.02]	[1.78–1.9]	[2.09–2.26]	[1.51–1.63]	[1.49–1.62]	[1.5–1.74]
Father or mother	2.22	2.11	2.38	1.92	1.87	2.1
	[2.15–2.3]	[2.03–2.21]	[2.26–2.51]	[1.79–2.07]	[1.73–2.03]	[1.82–2.43]
Father and mother	2.06	1.96	2.20	1.60	1.57	1.69
	[2.01–2.11]	[1.9–2.01]	[2.12–2.28]	[1.55–1.66]	[1.51–1.64]	[1.58–1.8]

The RRs were also adjusted for the following patient related confounders: other parent FH, age, sex, marital status, area of living (periphery rank, urban/rural, geographic administrative code), body mass index (BMI), and presence of hypertension (yes/no).

[a] The ARRs for diabetes were 1.51 [1.29–1.78] for paternal FH and 1.75 [1.49–2.06] for maternal FH, the ARRs for ASCVD were 1.50 [1.11–2.02] for paternal FH and 1.14 [0.9–1.43] for maternal FH.

For example, the RR that a subject would develop diabetes before the age of 40 years varied between 5.84 (95% CI 5.19–6.56), 2.52 (95% CI 2.12–2.99, $P < .001$), and 1.90 (95% CI 1.45–2.50, $P = .08$), depending on whether both parents were diagnosed with diabetes before the age of 60 years, at the age of 61–70 years, or at the age of 71–80 years, respectively. Similarly, a history of both parents having diabetes before the age of 60 years was associated with a RR of 5.84 (95% CI 5.19–6.56), 2.85 (95% CI 2.62–3.10, $P < .001$), and 1.05 (95% CI 0.92–1.21, $P < .001$) that the subject would develop diabetes before the age of 40 years, at the age of 41–50 years, or 51–60 years, respectively (Supplementary Table S2 in Supplementary Appendix S3). This analysis also included subjects excluded from the prediction model who had met the case definition prior to the index date of January 1, 2010.

A difference was also noted between the RR of paternal and maternal history of disease depending on the sex of the subject. In the same example above (parent diagnosed before 60, subject diagnosed before 40), a history of a father with diabetes was associated with a RR of 3.41 (95% CI 3.05–3.82) that his son would develop diabetes, versus 3.06 (95% CI 2. 67–3.52) for his daughter ($P = .2$). Likewise, a history of a mother with diabetes was associated with a RR of 4.21 (95% CI 3.73–4.75) that her daughter would develop diabetes, versus 3.26 (95% CI 2.94–3.63) for her son ($P = .002$, Supplementary Table S2 in Supplementary Appendix S3). As can be seen, these differences were only significant for maternal history and should also be interpreted with caution as we found a significant interaction between paternal and maternal histories of disease (ANOVA, $P < .001$).

Diabetes—predictive modeling

For the 3 age groups of 40–55 years, using only the subject's own information, the model achieved an AUC of 0.77–0.78 for the prediction of new onset of diabetes within 10 years. The addition of parent information significantly improved the model's performance ($P < .005$), leading to an AUC of 0.79–0.80. The sensitivity increased by 5% for the 40–45 years age group and by only 1% for the 55–60 years age group. Considering 10% and 18% prevalence rates in these age groups, respectively, the number needed to screen (ie, number of patients needed to have FH incorporated into their EHR to identify one otherwise undetected case) would be 200 for the 40–45 years age group and 550 for the 55–60 years age group. Further details are found in Table 3 and Figure 2. The

Table 3. Comparison of model performance (AUC), with and without parent information

Patient age	Patient data only	Patient+parents data (case definition only)		Patient+parents data (full EHR)	
	AUC [95% CI]	AUC [95% CI]	*P* value	AUC [95% CI]	*P* value
Diabetes					
40	0.77 [0.75–0.79]	0.79 [0.77–0.81]	.002[a]	0.79 [0.77–0.81]	.02[a]
45	0.78 [0.76–0.8]	0.80 [0.78–0.82]	<.001	0.80 [0.78–0.82]	<.001
50	0.77 [0.76–0.79]	0.79 [0.77–0.81]	<.001	0.79 [0.77–0.81]	<.001
55	0.79 [0.76–0.81]	0.79 [0.77–0.82]	.299	0.79 [0.77–0.82]	.518
All	0.79 [0.77–0.79]	0.80 [0.79–0.81]	<.001	0.80 [0.79–0.81]	<.001
CVD					
40	0.74 [0.71–0.77]	0.74 [0.71–0.78]	.227	0.74 [0.71–0.77]	.986
45	0.73 [0.71–0.76]	0.74 [0.71–0.77]	.049	0.74 [0.72–0.77]	.017
50	0.73 [0.7–0.75]	0.73 [0.7–0.75]	.836	0.73 [0.7–0.75]	.691
55	0.71 [0.67–0.75]	0.70 [0.67–0.74]	.251	0.69 [0.66–0.73]	.015
All	0.74 [0.73–0.76]	0.74 [0.73–0.76]	.15	0.74 [0.73–0.76]	.776

Note: Each model was compared to a baseline model derived from patient-only information. 95% CI and *P* values were calculated using DeLong's method.

[a] *P*-values are for the comparison with the "patient data only" model.

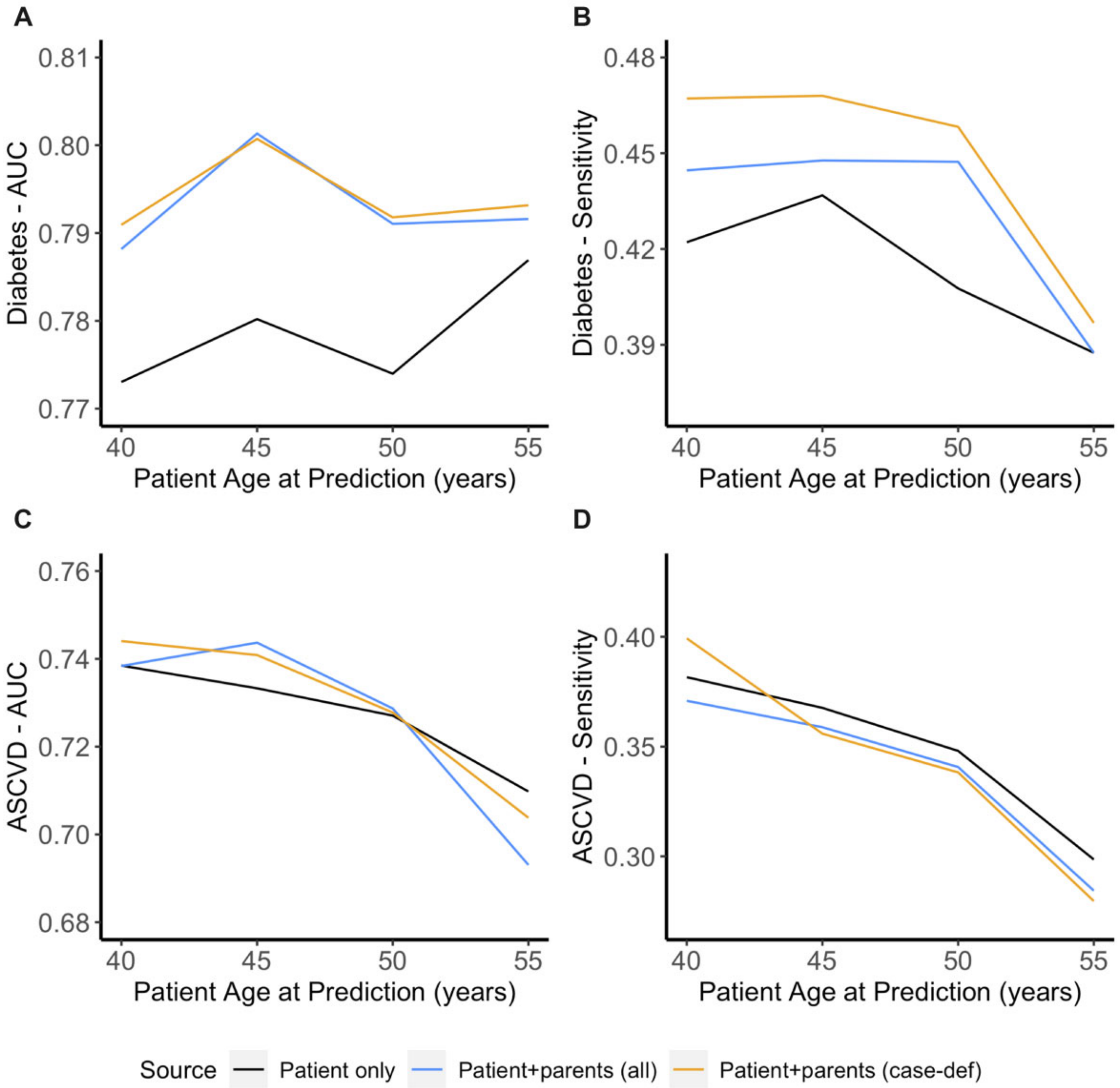


Figure 2. AUC (A + C) and sensitivity (B + D) of the prediction models, as a function of the age of the subject and the data used. Top (A + B): plots for the diabetes prediction models; Bottom (C + D): plots for the ASCVD prediction model. The added value of family history (blue and orange lines) over patient only information (black line) is shown.

Table 4. Top 20 risk factors for developing diabetes within the next 10 years[a]

Child/subject	Father	Mother
Lab: glucose (5.56–6.94)	DX: Diabetes mellitus	DX: Diabetes mellitus
Lab: hemoglobin A1C (39–47)	Lab: Hemoglobin A1C (H)	Lab: Hemoglobin A1C (H)
Lab: triglycerides (H)	DX: Diabetic retinal microaneurysms	Meds: blood glucose lowering agents, excluding insulin
Dem: Jewish	Meds: blood glucose lowering agents, excluding insulin.	Meds: Insulin and analogues
Dem: district	Lab: MCV (L)	DX: Diabetes mellitus, adult onset
Dem: sex	Lab: Microcytes % (H)	Lab: Iron (L)
Lab: uric acid (H)	DX: Diabetes mellitus, adult onset	DX: Diabetic retinal microaneurysms
Dem: number of children	DX: Retinopathy diabetic	Proc: Panoramic X-ray
Lab: GPT (H)	Meds: Insulin and analogues	Meds: High ceiling diuretics
DX: smoker	Lab: Glucose (H)	DX: Retinopathy, diabetic
Lab: HDL cholesterol (L)	Lab: Phosphorus (H)	Lab: CPK MB (taken)
DX: obesity	Lab: Fructosamine (H)	Lab: Album/creatinine ratio (H)
DX: essential hypertension	Lab: MCH (L)	DX: Biliary colic
DX: hyperglycemia	Lab: Hypochromia (L)	Lab: Bilirubin indirect (L)
DX: impaired fasting glucose	Lab: Albumin (L)	Lab: Microalbumin urine, (H)
Meds: ACE inhibitors	Lab: HCO3 (H)	DX: Congestive heart failure
DX: obesity (BMI > 30)	DX: Peripheral vascular disease	Lab: Albumin (L)
Lab: LDL cholesterol (taken)	Proc: Panoramic X-ray	Lab: Fructosamine (N)
Lab: WBC (H)	DX: Diabetes, routine follow-up	DX: Dehydration
DX: fatty liver	Proc: Hospitalization	Lab: Microalbumin, 24 h urine (taken)

[a] Variables are ordered by their relative contribution (gain) for each tree in the model. Lab tests: unless a specific threshold/value is indicated, these refer only to the ordering of the test. H: high; L: low; Dem: demographics; DX: diagnosis; Meds: medications; Lab: laboratory. The xgb.importance function in R was used to extract the relative contribution (gain) of each feature for each tree in the model, summed over all trees.

performance metrics for all models and all age groups are provided in Supplementary Table S3 in Supplementary Appendix S3. Among all age groups studied, maternal information was more predictive than paternal information for future diabetes in the offspring (Supplementary Figure S1 in Supplementary Appendix S3).

For the subject, the top risk factors for diabetes by model information gain were laboratory tests suggestive of prediabetes hyperglycemia (glucose levels of 5.56–6.94 mmol/L [100–125 mg/dL] or hemoglobin A1C of 39–47 mmol/mol [5.7%–6.4%]). Demographic features such as male gender, number of children, place of residence, and ethnicity were also associated with an increased risk for developing diabetes. In addition, features suggestive of metabolic syndrome were predictive of diabetes (hypercholesterolemia, hyperlipidemia, obesity, hypertension). For both parents, the top risk factors for future diabetes in their child were all related to the diagnosis of diabetes and its complications recorded in the parental EHR. A detailed summary is found in Table 4.

ASCVD

The subjects included 11 715 cases of ASCVD (10.6%). The median age of diagnosis (as recorded in the EHR) was 51.0% (IQR 46.5–54.5) and 72.2% were male ($OR = 3.0$, Table 1). The ARR for a subject having ASCVD was 1.50 (95% CI 1.11–2.02) if the father had CVD and 1.14 (95% CI 0.9–1.43) if the mother had CVD (Table 2). Overall, the addition of parents' information provided marginal value for the prediction of ASCVD. Other than the age group of 45–50 years, in which parents' information provided small but statistically significant improvement to the model's accuracy (AUC of 0.744 vs 0.733, $P = .017$), in all other age groups no such improvement was noted. Nonetheless, cases where a high risk for ASCVD was predicted based on the parents' data, were indeed at increased risk for ASCVD, regardless of the outcome predicted based on the patient's own EHR (Supplementary Table S2 in Supplementary Appendix S3). Further information on the ASCVD model and analysis can be found in the Supplementary Material (Supplementary Appendix S4).

DISCUSSION

Parental information from data-driven family history (DDFH) provided some improvement for the prediction of diabetes risk ($AUC = 0.8$ vs 0.78, $P < .001$), especially among subjects younger than 55 years of age. For ASCVD, while positive FH was associated with increased risk (RR of up to 3.11), it did not provide significant predictive value beyond the subject's own information when incorporated within a prediction model. We found that maternal medical history was more predictive of the subject's health than paternal medical history (Supplementary Figure S1 in Supplementary Appendix S3 and Supplementary Figure S1 in Supplementary Appendix S4). We found no significant improvement when including parents' entire detailed EHR, compared to using binary flags indicating whether each parent met the diabetes case definition.

There are several potential practice implications for adopting DDFH. In current clinical practice, FH is based on patients' self-reports. These self-reports are often unavailable at the point of care as they are either missing altogether, obtained too late (ie, after a patient has been diagnosed), or with inaccessible documentation (eg, hidden as free-text within a clinical note).[1,17,20] DDFH can make this information more easily available and in a timely fashion. We show that the number of patients needed to be screened to identify one otherwise missed subject would be 550 for the 55–60 years age group, 200 for the 40–45 years age group, and possibly even lower for younger patients. Furthermore, we examine the added value of DDFH in comparison to a RandomForest machine learning model that takes into account the *entire* EHR record of the patient. In reality, such models often do not exist, thus this may represent an overly ambitious bar to surpass. Testing the contribution of DDFH for overall risk stratification, we show that its RR is around 2.0 for diabetes and 1.5 for ASCVD, making it a significant risk factor to consider when evaluating a patient at the clinic.

From a research standpoint, the current study provides a unique opportunity to examine the contribution of FH for specific health conditions. Traditionally, studies aiming to evaluate the contribution of FH to the prediction of a subject's health outcomes are based on self-reported FH provided by the subject. These studies often suffer from several limitations intrinsic to the design, including the lack of a standard case definition for what constitutes a positive FH, inability to validate the subject's self-reported FH, and biases in reporting such as recall bias (eg, patients with diabetes are more likely to know if their relatives had diabetes than patients without diabetes). These limitations can lead to both under- and overestimation of risk associated with a positive FH. DDFH can overcome many of these challenges.

Deriving DDFH requires linkage of records between family members, who often are not part of the same care provider or insurer, and a framework for obtaining parents' consent to use their information to predict the clinical outcome of their children. However, our results suggest that full linkage of the records may not be necessary, and that sharing only EHR-derived case definitions between parents and their children may be sufficient for improving clinical predictions. Interoperability will be an important challenge for adoption of this platform, perhaps making it initially more relevant for large care providers with many family groups in the covered population. If proven successful, solutions such as Fast Healthcare Interoperability Resources (FHIR) or interoperability platforms such as those used for opiate prescriptions, may help introduce this platform into smaller healthcare systems. Implementation of DDFH also raises certain ethical questions that will need to be addressed. Data availability can differ among subjects, as not all patients live within the same health system catchment area as their parents. These variations in data availability could affect prediction accuracy and, subsequently, the quality of care received by different patients. Presently, healthcare decision-making does not sufficiently incorporate family history. While this study does not propose a comprehensive solution to address this problem, it demonstrates that, when accessible, DDFH information can be used to improve upon current predictions. To ultimately extend this opportunity to all, a long-term goal of this study is to build the evidence base for the value of data-driven FH for clinical decision-making. Such evidence is necessary to motivate the design and implementation of a future opt-in data sharing framework, enabling individuals to voluntarily share certain aspects of their medical history with family members, including across healthcare systems.

The results of our data-driven approach are in-line with those found in previous studies. Prior studies have failed to show that FH has a significant contribution for predicting ASCVD over and above the information available on the subject and it is not part of accepted clinical risk scores (eg, Framingham risk score, the pooled cohort equation, or the European SCORE).[38] Our finding that maternal medical history has greater predictive value than paternal medical history for the prediction of diabetes is also supported by previous studies.[39–42] Our data-driven design and the large sample size of the study population also enable us to answer questions that were not feasible in previous studies. For example, we can observe that the added predictive value contributed by parental information declines with increasing age of the subject. There are several possible explanations for this observation: younger subjects have less available information, making parental data more crucial; the increasing impact of environmental factors and exposures over time (eg, stress, obesity, etc.); and earlier predictions correlating with younger parent age, which suggests an earlier diagnosis for the parent and increased risk for the subject.[43]

This study has several limitations. First, while data records reach back a few decades in some cases, in many cases comprehensive digital documentation was only available from 2003 onward. Thus, parental information, in some cases dating back to earlier years, may be incomplete. While these data are often captured indirectly in future years for which we have documentation (eg, purchase of insulin anytime between 2003 and the index date for parents with an earlier diagnosis of diabetes), exact dates prior to 2003 were not always available. This would be less of an issue in the future, as with time, more data accumulate in the electronic medical records of subjects and their parents, potentially improving such predictions. Second, to preserve patient privacy, we did not have access to the full medical charts of the subjects. Thus, we were not able to validate our case definitions. Nevertheless, we used case definitions based on prior studies conducted using the same data repository and validated ICD-based case definitions that are commonly used for such purposes. Third, as we did not have information on the setting in which lab tests were taken, we could not rely on lab tests alone to make the diagnosis of diabetes. Thus, while 2 separate measures of blood glucose over 11.11 mmol/L (200 mg/dL) may be sufficient for the diagnosis of diabetes in a clinical setting, we had to exclude these patients if there was no other documentation to support this diagnosis.

Finally, the data used in the present study, which includes the information necessary to link medical records between family members, are not readily available in many healthcare systems. Thus, implementation of the models as-is may not presently be possible in other settings. Nevertheless, as mentioned above, full linkage of the records may not be necessary, as sharing of an EHR-derived binary case definition may suffice. Previous studies have demonstrated creative ways of linking EHR records without explicit linkage data but instead using patients' contact information.[44] Our findings and the framework we provide can be useful for a cost-effectiveness analysis aimed at quantifying the tradeoff between the information gain and the risk it poses for patient privacy and the financial costs of creating a family-linked EHR system. This analysis can inform scientists and decision makers considering the development and adoption of a robust policy framework to facilitate safe information sharing among family members and their care providers.

In conclusion, data from the electronic medical records of family members have the potential to address many of the shortcomings of relying on patient self-reports to collect family history information. We show that parental information can significantly improve clinical predictions for some conditions, but not for all, and that this information is especially useful among younger adults. Enriching patients' medical records with EHR-driven FH can improve screening and delivery of personalized medicine. Future studies can build upon the framework provided in this study to investigate the effect of maternal and paternal health history across a broad range of diseases.

COPYRIGHT ASSIGNMENT

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide license to the Publishers and its licensees in perpetuity, in all forms, formats, and media (whether known now or created in the future), to (1) publish, reproduce, distribute, display, and store the Contribution, (2) translate the Contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or, abstracts of the Contribution, (3) create any other derivative work(s) based on the Contribution, (4) to exploit all subsidiary rights in the Contribution, (5) the inclusion of electronic links from the Contribution to third party material where-ever it may be located; and, (6) license any third party to do any or all of the above.

PATIENT INVOLVEMENT

Patients were not involved in the design or conduct of this study.

TRANSPARENCY DECLARATION

BYR affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned (and, if relevant, registered) have been explained.

FUNDING

This research was funded by grant number R01-HG009129 from the NIH. The funder played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

AUTHOR CONTRIBUTIONS

All authors took part in planning the research. YB-C, DT, and DK analyzed data. IG prepared data for analysis. BYR, NB, DS, and ML-R supervised the research. All authors worked on interpreting the results. All authors reviewed the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. BYR and YB-C are responsible for the integrity of the work as a whole.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/disclosure-of-interest/ and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work. DT currently works at Mobileye.

DATA AVAILABILITY

The data analyzed in this study may not be shared due to medical privacy regulations.

REFERENCES

1. Guttmacher AE, Collins FS, Carmona RH. The family history – more important than ever. *N Engl J Med* 2004; 351 (22): 2333–6.
2. Fuller M, Myers M, Webb T, Tabangin M, Prows C. Primary care providers' responses to patient-generated family history. *J Genet Couns* 2010; 19 (1): 84–96.
3. Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med* 2002; 4 (4): 304–10.
4. Doerr M, Edelman E, Gabitzsch E, Eng C, Teng K. Formative evaluation of clinician experience with integrating family history-based clinical decision support into clinical practice. *J Pers Med* 2014; 4 (2): 115–36.
5. NHGRI. *Family History Implementation in the Challenging Setting of Routine Clinical Care*. National Advisory Council for Human Genome Research; 2012. https://www.genome.gov/Pages/About/NACHGR/Sept2012AgendaDocuments/ConceptClearance_FamilyHistory_AWise.pdf.
6. Williams RR, Hunt SC, Heiss G, *et al.* Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (the Health Family Tree Study and the NHLBI Family Heart Study). *Am J Cardiol* 2001; 87 (2): 129–35.
7. Veronesi G, Gianfagna F, Giampaoli S, *et al.* Improving long-term prediction of first cardiovascular event: the contribution of family history of coronary heart disease and social status. *Prev Med* 2014; 64: 75–80.
8. Lloyd-Jones DM, Nam B-H, Ralph B, *et al.* Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults: a prospective study of parents and offspring. *JAMA* 2004; 291 (18): 2204–11.
9. Vernon SW. Risk perception and risk communication for cancer screening behaviors: a review. *J Natl Cancer Inst Monogr* 1999; 1999 (25): 101–19.
10. McDowell ME, Occhipinti S, Chambers SK. The influence of family history on cognitive heuristics, risk perceptions, and prostate cancer screening behavior. *Health Psychol* 2013; 32 (11): 1158–69.
11. Harrison TA, Hindorff LA, Kim H, *et al.* Family history of diabetes as a potential public health tool. *Am J Prev Med* 2003; 24 (2): 152–9.
12. Volk LA, Staroselsky M, Newmark LP, *et al.* Do physicians take action on high risk family history information provided by patients outside of a clinic visit? *Stud Health Technol Inform* 2007; 129 (Pt 1): 13–7.
13. Gottesman O, Kuivaniemi H, Tromp G, *et al.*; eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013; 15 (10): 761–71.
14. Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011; 3 (79): 79re1.
15. Kerber RA, Slattery ML. Comparison of self-reported and database-linked family history of cancer data in a case-control study. *Am J Epidemiol* 1997; 146 (3): 244–8.
16. Ziogas A, Anton-Culver H. Validation of family history data in cancer family registries. *Am J Prev Med* 2003; 24 (2): 190–8.

17. Benson L, Baer HJ, Greco PJ, Kaelber DC. When is family history obtained? – lack of timely documentation of family history among overweight and hypertensive paediatric patients. *J Paediatr Child Health* 2010; 46 (10): 600–5.
18. Janssens ACJW, Henneman L, Detmar SB, *et al.* Accuracy of self-reported family history is strongly influenced by the accuracy of self-reported personal health status of relatives. *J Clin Epidemiol* 2012; 65 (1): 82–9.
19. Giovanni MA, Murray MF. The application of computer-based tools in obtaining the genetic family history. *Curr Protoc Hum Genet* 2010; Chapter 9: Unit 9.21.
20. de Hoog CLMM, Portegijs PJM, Stoffers HEJH. Family history tools for primary care are not ready yet to be implemented. A systematic review. *Eur J Gen Pract* 2014; 20 (2): 125–33.
21. Yoon PW, Scheuner MT, Jorgensen C, Khoury MJ. Developing family healthware, a family history screening tool to prevent common chronic diseases. *Prev Chronic Dis* 2008; 6 (1): A33.
22. Rich EC, Burke W, Heaton CJ, *et al.* Reconsidering the family history in primary care. *J Gen Intern Med* 2004; 19 (3): 273–80.
23. Acheson LS, Wiesner GL, Zyzanski SJ, Goodwin MA, Stange KC. Family history-taking in community family practice: Implications for genetic screening. *Genet Med* 2000; 2 (3): 180–5.
24. Reid G, Emery J. Chronic disease prevention in general practice – applying the family history. *Aust Fam Physician* 2006; 35 (11): 879–82. 884–5
25. Leventer-Roberts M, Gofer I, Barak Corren Y, *et al.* Constructing data-derived family histories using electronic health records from a single healthcare delivery system. *Eur J Public Health* 2020; 30 (2): 212–8.
26. Estruch R, Ros E, Salas-Salvadó J, *et al.*; PREDIMED Study Investigators. Primary prevention of cardiovascular disease with a mediterranean diet supplemented with extra-virgin olive oil or nuts. *N Engl J Med* 2018; 378 (25): e34.
27. Shmueli A. Switching sickness funds in Israel: adverse selection or risk selection? Some insights from the analysis of the relative costs of switchers. *Health Policy* 2011; 102 (2–3): 247–54.
28. Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015; 13 (1): 1.
29. Ward A, Sarraju A, Chung S, *et al.* Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *NPJ Digit Med* 2020; 3 (1): 125–7.
30. Goff DC, Lloyd-Jones DM, Bennett G, *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. *J Am Coll Cardiol* 2014; 63 (25 Pt B): 2935–59.
31. Khokhar B, Jette N, Metcalfe A, *et al.* Systematic review of validated case definitions for diabetes in ICD-9-coded and ICD-10-coded data in adult populations. *BMJ Open* 2016; 6 (8): e009952.
32. Elley CR, Robinson E, Kenealy T, Bramley D, Drury PL. Derivation and validation of a new cardiovascular risk score for people with type 2 diabetes: the new zealand diabetes cohort study. *Diabetes Care* 2010; 33 (6): 1347–52.
33. Barak-Corren Y, Chaudhari P, Perniciaro J, *et al.* Prediction across healthcare settings: a case study in predicting emergency department disposition. *NPJ Digit Med* 2021; 4 (1): 169.
34. Li W, Song Y, Chen K, *et al.* Predictive model and risk analysis for diabetic retinopathy using machine learning: a retrospective cohort study in China. *BMJ Open* 2021; 11 (11): e050989.
35. Sarraju A, Ward A, Chung S, *et al.* Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart* 2021; 8 (2): e001802.
36. Robin X, Turck N, Hainard A, *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12: 77.
37. 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 (3): 837–45.
38. Bittencourt MS. Family history of cardiovascular disease: how detailed should it be? *Mayo Clin Proc* 2018; 93 (9): 1167–8.
39. Bener A, Yousafzai MT, Al-Hamaq AO, Mohammad A-G, DeFronzo RA. Parental transmission of type 2 diabetes mellitus in a highly endogamous population. *World J Diabetes* 2013; 4 (2): 40–6.
40. Papazafiropoulou AK, Papanas N, Melidonis A, Maltezos E. Family history of type 2 diabetes: does having a diabetic parent increase the risk? *Curr Diabetes Rev* 2017; 13 (1): 19–25.
41. Karter AJ, Rowell SE, Ackerson LM, *et al.* Excess maternal transmission of type 2 diabetes. The Northern California Kaiser Permanente Diabetes Registry. *Diabetes Care* 1999; 22 (6): 938–43.
42. Prasad RB, Lessmark A, Almgren P, *et al.* Excess maternal transmission of variants in the THADA gene to offspring with type 2 diabetes. *Diabetologia* 2016; 59 (8): 1702–13.
43. Allport SA, Kikah N, Saif NA, Ekokobe F, Atem FD. Parental age of onset of cardiovascular disease as a predictor for offspring age of onset of cardiovascular disease. *PLoS One* 2016; 11 (12): e0163334.
44. Polubriaginof FCG, Vanguri R, Quinnies K, *et al.* Disease heritability inferred from familial relationships reported in medical records. *Cell* 2018; 173 (7): 1692–704.e11.