HHS Public Access

Author manuscript

Nat Health. Author manuscript; available in PMC 2026 August 03.

Published in final edited form as:

Nat Health. 2026 March ; 1(3): 316–325. doi:10.1038/s44360-026-00054-9.

Access to care affects electronic health record reliability and AI-driven disease prediction

Anna Zink[1], **Hongzhou Luan**[2], **Irene Y. Chen**[3]

[1]Tufts University, Medford, MA, USA

[2]University of California, Berkeley, CA, USA

[3]University of California, Berkeley, CA, USA

Abstract

Despite well-documented healthcare access disparities, their impact on electronic health record (EHR) reliability and resulting clinical prediction models remains poorly understood. Analyzing 205,186 participants from the All of Us Research Program, we found that participants with cost-constrained or delayed care had worse EHR reliability for 73% of examined medical conditions as measured by participant self-reported conditions, driven in part by lower visit rates. In a type 2 diabetes prediction task, including participant self-reported conditions significantly improved the predictive performance for participants with lower access to care and improved targeting of low access patients who would go on to develop type 2 diabetes. This study demonstrates that healthcare access systematically affects both data quality and clinical prediction performance, suggesting that improving health equity requires addressing both data collection biases and algorithmic limitations. Our findings provide an empirical foundation for developing clinical prediction systems that work effectively for all patients regardless of barriers to access.

Introduction

The ability to access healthcare fundamentally shapes health outcomes, yet persistent social and economic barriers keep essential care out of reach for many patients. Disparities in access to care have been documented for many patient groups across the world, including patients from lower income populations, those with lower educational attainment, and patients that identify as Black and Hispanic or Latino patients, among others,[1–4] As healthcare systems increasingly leverage predictive models built on large administrative claims and electronic health record datasets to guide clinical decisions[5–8], this access disparity creates a concerning paradox: the patients most affected by access barriers may be systematically misrepresented in the data used to identify those in need of care.

Corresponding Author: Anna Zink, Tufts University, Medford, MA 02155, anna.zink@tufts.edu.

Author Contributions Statement

AZ and IC conceived of the study and study design, AZ and HZ analysed and interpreted the data, AZ wrote the paper, and AZ, HZ, and IC edited the paper. All authors read and approved the final paper.

Competing Interests Statement

IC consults for Flatiron Health. AZ has equity in Knit Health Technologies Inc. HZ declares no competing interests.

Despite extensive documentation of healthcare access disparities, their effects on EHR data and resulting clinical predictions remain largely unknown. This knowledge gap is particularly concerning as healthcare systems increasingly base clinical decisions on machine learning models trained using electronic health record data. Data quality issues in the EHR are a well-explored topic[9], with many studies documenting the effect of EHR quality on research results[10,11], but no studies to our knowledge document the impact of access on EHR data. Similarly, previous research has suggested that limited access reduces model performance but rely on indirect proxies like race or insurance status rather than direct measures of access.[12–16] Understanding how access shapes both data collection and algorithmic performance requires datasets that directly measure healthcare access and link it to clinical records.

Our study uses linked survey and electronic health data from the All of Us Research Program to quantify the relationship between health care access, EHR reliability, and clinical risk prediction (Figure 1a).[17] Developed by the National Institute of Health, the All of Us Research Program is one of the largest biomedical data resources in the United States. It recruits participants across the U.S., who provide data via survey but may also opt to link their EHR data from participating sites. It's linkage of self-reported access to care with EHR data provides a unique opportunity to study the implications of access to care on EHR reliability and EHR-based clinical risk prediction.

To study EHR reliability, we quantify the relationship between access to care and missing health condition data in the EHR record using self-reported health information collected upon enrollment in the All of Us Research Program. We assume that a participant's report of a health condition reflects either the true presence of the condition or a belief that they have the condition. In either case, the condition should plausibly appear in the EHR as a diagnosis or as documentation of diagnostic evaluation. We compare rates of EHR reliability across 37 self-reported health condition, stratified by access to care groups, to assess whether reliability differs for participants with lower access to care. We further explore frequency of visits to the health care system as a potential mechanism for differences in EHR reliability rates.

To evaluate the extent to which clinical machine learning methods may create biased predictions for participants with lower access to care, we focus on one clinical task: the prediction of (T2D) diabetes incidence over a two-year time horizon. Disease incidence is widely recognized as an area in which machine learning and predictive analytics have the potential to improve clinical targeting.[18–21] T2D is an especially valuable case study, given its clinical importance and potential for machine learning: while early intervention can greatly improve health outcomes, it remains a leading causes of death worldwide and the prevalence of T2D has been increasing in recent years.[22,23]

Results

We start with data from 633,540 All of Us participants that have enrolled and completed the initial steps of the program, which entails a preliminary survey on patient demographic information. After restricting the sample to participants that responded to the "Health Access

& Utilization" survey (N=305,857) in order to define access groups, our analytic sample includes 205,186 All of Us participants for which we had self-reported data on healthcare access and linked EHR data (Figure 1b). In Extended Data Table 1, we compare sample characteristics of the All of Us analytic sample, All of Us sample more broadly, and the larger U.S. (based on estimates from the 2023 American Community Survey). All of Us participants are more likely to be older, female, low-income, and covered by Medicare and/or Medicaid compared to the general U.S. population. Restricting to the All of Us analytic sample, increases the share of sample participants that are female, white, higher-income, and have insurance through their employer.

More than half of the 205,186 participants indicated that they had trouble accessing care, either due to cost-constrained or delayed care in the previous year: 87,921 participants (42.8%) indicated trouble affording care in the past year, 73,445 participants (35.8%) indicated delaying care, and 50,490 participants (24.6%) indicated both. We defined participants as having cost-constrained care and/or delayed care based on survey responses to the "Health Access & Utilization" survey, which asked participants whether they had trouble affording care or had delayed care in the past 12 months. Those with cost-constrained care indicated problems paying for prescriptions, dental care, and specialty care, among other things. Common reasons for delaying care included nervousness or having to pay out of pocket (Extended Data Figure 1). Participants that did not fall into either the cost constrained and/or delayed care group were classified as having standard care. We further defined levels of access to care based on the number of questions in which participants indicated they had trouble affording care or had delayed care (see Materials & Methods for more information on defining access groups).

Table 1 compares participants included in the analytic sample by access groups. Participants with cost-constrained and/or delayed care were more likely to have Medicaid insurance, be uninsured, and have lower annual household incomes. They were also more likely to be younger, female, Black, and Hispanic (Table 1). Participants with cost-constrained and/or delayed care reported higher rates for each of the ten most common self-reported conditions compared to those with standard care (Extended Data Table 2). Prevalence rates were age-adjusted to account for differences in age between the two groups. For example, participants with cost-constrained and/or delayed care had higher age-adjusted rates of depression (46.0 versus 30.2) and asthma (27.4 versus 19.8). We used a logistic regression to assess the statistical significance of the relationship between access and self-reported condition, adjusting for age. P-values from the coefficient on access variable for each regression were adjusted for multiple comparisons based on Benjamini-Hochberg procedure. All differences were statistically significant.

Access to Care and EHR reliability

To understand how access to care impacts EHR reliability, we compare self-reported versus EHR-reported data on 37 health conditions. Self-reported health conditions are constructed based on participant survey response to whether a participant has or has every had a specific condition. EHR health conditions are constructed using condition concepts. Supplementary Table 1 includes information on how each of the 37 conditions were defined in both the

EHR and survey. There are two margins of discordance between these values. First, a participant might say they have had a condition, but no record exists in the EHR. Second, a participant might say they have never had a condition, but there is a record of it in the EHR. The former discrepancy may indicate gaps in medical documentation, care access, or errors in self-reporting, while the latter discrepancy may reflect clinical assessments that differ from participant self-perception. We assume that a condition reported in the EHR and by self-report is the true condition status; however, a condition that is not reported in the EHR or by self-report may not reflect the true condition status if the participant has an undiagnosed condition. We are therefore only able to evaluate EHR data quality with respect to diagnosed conditions. Assuming no difference in self-reporting error across groups, this approach measures potential care gaps and documentation failures that may disproportionately affect those with access barriers, providing a clearer measurement of systematic EHR reliability differences.

To assess whether EHR data may be differentially missing health information for those with lower access to care, we focus our analysis on conditions that the participant says they have or have had. We calculate the EHR reliability rate as the percent of participants that report having a condition that also have that condition documented in the EHR records for each access group (standard care, delayed care, and cost-constrained care). When we compare the EHR reliability of 37 conditions for each access group using a two sample proportions z-test, we find that participants with cost-constrained or delayed care had statistically significant lower rates of EHR reliability for 27 out of 37 conditions compared to the standard care group (Figure 2), using the Benjamini-Hochberg procedure to control for multiple comparisons. For example, the EHR reliability rate for hypertension was 82.2% for those with standard care versus 78.2% and 76.9% for cost-constrained care and delayed care, respectively. In other words, among participants that report having hypertension, we find a corresponding record of hypertension in the EHR data for 82.2% of participants with standard care versus 76.9% of participants with delayed care.

If the assumption of no difference in misreporting rates by access group does not hold, it is possible that the lower EHR reliability rates we observe are in part due to higher misreport by patients with lower access to care; false reports of a condition will artificially inflate the denominator in the calculation of the EHR reliability rate. To test this assumption, we consider the second margin of error: instances where the participant does not self-report a condition, but the EHR records the condition (as of the point in time when the survey was answered). We calculate the percent of participants that have a condition reported in the EHR that also self-report that condition, for each access group. This measure is similar to the EHR reliability measure, but we take the EHR as the validating source rather than the self-report data. If those with lower access to care have higher rate of self-reporting error, we might expect that they would have lower rates of self-report reliability. However, we find that the self-report reliability measure is higher for participants with lower access to care on 32 of the 37 conditions (24 of which are statistically significant based on a two sample proportions z-test, controlling for multiple comparisons with the Benjamini-Hochberg procedure). While these findings cannot definitively rule out misreporting error as a confounder in the differences in EHR reliability, they suggest that those with low access to

care may have lower rates of misreport than those with high access which would suggest that the gaps in EHR reliability we observe may be more conservatively estimated.

Access to care likely influences the visit frequency of healthcare visits. Figure 3 shows the cumulative rates of time to first visit by access group over a two-year observation window, starting at the participant's All of Us enrollment date. We find that incidence of time to first-visit is significantly higher for participants in the standard care group compared to those in the cost-constrained and/or delayed care group at every month over the two-year time horizon. The gap between the two groups remains consistent at about 5%, indicating that at least some of the lower reliability in the EHR may be partially explained by visit frequency. To assess this mechanism, we compare nested linear regression models with and without adjustment for visit frequency. Among participants that self-report a condition, the probability that the condition will be recorded in the EHR increases by 2.9 percentage points (P Value <0.001) for someone with standard care compared to someone without standard care, controlling for condition type (i.e., an indicator for whether the participant self-reported a condition for each of the 37 health conditions examined). When we additionally control for visit intensity, measured as the number of health system visits recorded for each participant, we find that that this probability decreases from 2.9% to 1.1%, suggesting that differential rates of health system visits mediates part of the relationship between EHR reliability and access to care. The full regression specification can be found in the Materials and Methods section.

Implications for Clinical Prediction Models

When EHR data are used for predictive algorithms, differences in medical condition reliability indicate that participants with lower access to care may have worse quality feature data or outcome labels. To illustrate this, we develop and evaluate a prediction task of T2D incidence. T2D had the highest EHR reliability among the set of health conditions considered in our EHR reliability analysis (Figure 2), ensuring a more reliable outcome label for prediction. However, many of the features used for analysis may suffer from lower EHR reliability.

We test for this by developing a T2D prediction model and comparing the predictive performance of the model across access groups. We designed the prediction model to align with other EHR-based prediction models for T2D.[26,27] The training sample included 93,593 participants without a record of T2D in the EHR at the time they completed their All of Us baseline survey. T2D was identified using OMOP condition concepts and lab data on HbA1c and glucose levels. We considered someone diagnosed with T2D if they had a T2D condition code, HbA1c >= 6.5%, fasting glucose >= 126, or random glucose >= 200.

We evaluated four specifications for building the T2D prediction model: (1) a baseline model that is set up from the perspective of the health system and uses only data available in the EHR, (2) a baseline + access model that incorporates indicators for cost-constrained and delayed care as features in the clinical prediction model, (3) a baseline + SES model that incorporates feature data correlated with healthcare access including insurance coverage, socioeconomic data, and prior visit information, and (4) a baseline+self-report model that adds self-reported condition data as features in the model.

Incorporating access features – either directly or through proxy variables such as insurance and SES – could improve the predictive performance for those with cost-constrained or delayed care if there was differential risk of T2D incidence for those with low access to care which was not captured by the existing features in the model.[30] In the baseline+access model we include access features directly; however health systems do not typically have a way to measure access directly. Therefore, in the baseline+SES model we incorporate data that is correlated with health care access, such as insurance, income, and prior visit history, and measured by health systems. Our final approach, the baseline+self-report model aims to improve the quality of the feature data used in the prediction model. In this example, including self-reported conditions as features could improve predictive performance by providing information to the model that was not available in the EHR features.

The outcome was 2-year T2D incidence. Baseline features included common risk factors for T2D such as obesity, high blood pressure, high triglycerides, demographic information recorded at the time of the index date (age, gender, race, and ethnicity), and additional clinical information from the EHR including recorded health conditions, lab results, clinical observations like blood pressure, and medication history.[18,24,28] All features except for age were constructed as indicator variables, with a value of 1 indicating the presence of a condition, record of medication, or membership in a group, and a value of 0 representing the lack of condition, record, or membership. We used a 2-year lookback window to construct the feature data. Following the literature on other prediction models using EHR and claims data, missing data were set to zero. There was no missing data on age. There were 494 features in total. The inclusion of both clinically validated risk factors and additional clinical features is common practice when developing a machine learning model. For more information on the feature data included in each model, feature construction, and model training can be found in the Materials and Methods section.

Among the 93,593 participants without a record of type 2 diabetes in the EHR at the time they completed their All of Us baseline survey, 2.9% of participants in the sample received a T2D diagnosis in the 2-year observation window. We note the same pattern of demographic differences between standard care, cost-constrained care, and delayed care access groups that we saw in the main analytic sample; namely, that participants with cost-constrained and delayed care were more likely to be younger, female, Black, and Hispanic compared to participants with standard care, and they were more likely to be covered by Medicaid, uninsured, and report lower annual household incomes (Extended Data Table 3). We recalculate EHR reliability rates among the diabetes sample across all conditions. We find higher EHR reliability rate for those with standard care than those with cost-constrained or delayed care (50.3% versus 47.7%, P Value < 0.001 based on a two sample proportions z-test), corresponding to a relative increase of 5.5% in EHR reliability for those with standard care. This difference includes variation in reliability within and across condition types.

The logistic regression model with L1 regularization or LASSO algorithm was used for model comparisons, as it outperformed the Ridge and random forests algorithms on the baseline model. The baseline model AUC was 0.76 (95% CI: 0.753 to 0.771). The AUC value for the standard care group was 0.76 (95% CI: 0.749 to 0.776), compared to 0.75 (95%

CI: 0.740 to 0.767) for the cost-constrained group and 0.77 (95% CI: 0.753 to 0.783) for the delayed care group. The difference in AUC between the standard care group and delayed care and cost-constrained care group were not statistically significant based on a P Value = 0.05, measured using DeLong's Method (Extended Data Table 4).[29]

The model correctly identified 64.2% (95% CI: 63.7 to 64.7) of new diabetes cases for the delayed care group and 66.9% (95% CI: 66.4 to 67.4) for the cost-constrained group, relative to 68.8% (95% CI: 68.4 to 69.3) for the standard care group. For participants with severe delayed care the sensitivity dropped even further to 63.4% (95% CI: 62.5 to 64.4). The specificity for the cost-constrained care and standard care group was similar at 70.6% (95% CI: 70.2 to 71.1) and 70.3% (95% CI: 69.9 to 70.7) whereas the delayed care group had a higher specificity of 74.7% (95% CI:74.2 to 75.1).

In Table 2, we compare AUC values for each group across the four specifications. AUC comparisons are calculated using DeLong's method and P-values are adjusted for multiple comparisons using the Benjamini-Hochberg procedure. The baseline+access model showed no improvements in AUC for any of the access groups. However, the baseline+SES and baseline+self-report models showed statistically significant increases in AUC values for all access groups. In Figure 4, we plot performance metrics for each method by access group. When adding SES-related variables there is an increase in sensitivity for both the delayed and cost-constrained group and only a small decrease in specificity. When adding patient self-reported conditions, the increase in sensitivity was even more marked, and there was a slightly larger decrease in specificity. Balanced accuracy greatly improved for the cost-constrained and delayed care groups when patient self-report data was added.

Higher sensitivity may lead to more targeted clinical interventions for cost-constrained and delayed care groups who ultimately develop diabetes. In this task, including self-reported data as features in the prediction model identified an additional 100 low-access participations who went on to develop T2D, a relative increase of 6.3%. However, this improvement came at the cost of 1,141 (or 2.3%) additional false positives (2.3% of low-access participants who did not develop T2D as recorded in our data). In contrast, among participants with standard care, adding self-reported data identified 28 additional true cases (or 2.4% increase) and 52 additional false positives (0.1%).

Race and access to care

Lower algorithmic performance for Black patients has been documented across numerous clinical settings.[13] Reduced performance could, at least partially, be explained by differences in access to care for Black patients.[16] We test this theory on our T2D sample. Black participants in the T2D sample report statistically significantly higher rates of delayed and cost-constrained care compared to White participants: 62.0% of Black participants say that they have delayed or cost-constrained care, compared to 50.5% of white participants. Black participants had statistically significant lower AUCs for the T2D task compared to White participants: the AUC value for Black participants was 0.71 (95% CI: 0.798 to 0.718) versus 0.76 (95% CI: 0.762 to 0.768) for White participants. When we stratify participants by access, the race gap in AUC largely remains. Among Black participants with standard care the AUC is 0.69 (95% CI: 0.675 to 0.709) versus 0.76 (95% CI: 0.759 to 0.768) for white

participants, and among Black participants without standard care the AUC is 0.70 (95% CI: 0.681 to 0.717) versus 0.77 (95% CI: 0.759 to 0.773) for white participants.

Discussion

The relationship between access to healthcare and clinical prediction has been relatively understudied to-date, despite the prevalence of access issues globally and an increasing reliance on prediction models to aid in clinical decision-making.[6] Using novel data from the All of Us Research Program, we find consistently lower EHR reliability rates across nearly thirty self-reported health conditions for those with cost-constrained or delayed care compared to those with standard care, confirming concerns related to the quality of EHR data for those with lower access to healthcare services. One reason we might observe lower rates of EHR reliability for those with lower access to care would be if participants with lower access to care were less likely to visit the doctor.[32] If participants with lower access to care do not visit the doctor or visit the doctor less frequently than those with standard care, information on their health status is less reliably collected in the EHR, resulting in higher rates of missing data for participants with lower access to care. Further analysis suggests that lower EHR reliability for those with lower access to care is at least partially explained by their lower visit rates to the health care system.

In a clinical prediction task of T2D incidence, we observed no difference in predictive performance for participants with low access compared to participants with standard care. However, when we incorporated SES and self-reported condition data into the T2D model, we significantly improved model performance for all participants but particularly for those with cost-constrained or delayed care. These improvements, most notably the inclusion of self-reported condition data, increased the sensitivity of the model, allowing for better identification of low access participants who would ultimately develop T2D, with only a relatively smaller drop in specificity.

Our findings reveal a critical blind spot in AI-driven healthcare: patients most in need of clinical intervention are systematically least visible to the algorithms design to help them. This ‘access-to-algorithm’ gap creates a dangerous feedback loop where limited healthcare access leads to sparse data, which produces less accurate predictions, potentially resulting in delayed care and worse outcomes for already vulnerable populations. This is concerning on two fronts. First, clinical risk prediction models are being adopted by an increasing number of health systems each year.[33] Second, structural barriers to care—such as insurance coverage gaps and rising healthcare costs—may be worsening, as recent reports highlight growing uninsurance rates and mounting affordability concerns.[34,35]

Addressing this issue will require both long-term policy change and immediate methodological interventions. Policy-based solutions—such as expanding insurance coverage or reducing out-of-pocket costs—represent the most direct path to improving access and, in turn, EHR data quality. Other policy solutions could focus on interoperability solutions that improve the sharing of data across systems. These solutions might be more feasible in settings with universal health care like the National Health Service. However, these changes take time to implement. In the interim, our findings suggest that integrating

alternative data sources, such as patient-reported health or alternative data collection methods like wearables, may help mitigate the limitations of EHR data for individuals with lower access to care. There is already growing momentum around incorporating patient-reported outcomes into predictive modeling.[36–38] In addition, methodological approaches such as imputing missing data, using proxy indicators of access (e.g., SDOH variables or clinical notes), and explicitly modeling visit frequency may improve model performance for underserved populations.[39,40] Finally, alongside technical solutions, governance and monitoring frameworks are essential to track and mitigate the clinical impact of reduced model performance for those with limited access to care.[41,42]

Our study extends prior research showing high rates of discordance between survey data and EHR data to focus on differential rates of discordance by access to care.[43,44] We further consider how differential rates of reliability impact downstream prediction tasks, which contributes to a growing body of work on the measurement and mitigation of potential sources of algorithmic bias in clinical prediction models.[45] While previous work has focused on bias with respect to protected classes such as those defined by race, ethnicity, or gender, here we focus on access as a potential mechanisms for reduced performance across groups. To date there has been a lack of empirical research quantifying the relationship between healthcare access and clinical prediction, in part due to data limitations in EHR and health claims data which do not record measures of healthcare access. Our findings provide quantitative evidence to support concerns around reduced prediction performance for those with lower access to care. However, we do not find that access explains performance gaps by race.

This work had several limitations. We relied on self-reported data to validate the reliability of EHR data. Self-reported data have inherent limitations and may not always be accurate.[46] Patients may underreport certain conditions if they are unfamiliar with the medical term or uncomfortable sharing certain medical information due to social stigma.[24,25] However, we argue that even if someone doesn't have the diagnosis, the act of reporting a condition still signals illness. In such case, the EHR's failure to capture this information means it inaccurately reflects participant's overall health status, which has downstream implications for prediction. Moreover, the focus of our analysis was on measuring group differences rather than group levels. To test whether observed differences in EHR reliability were potentially confounded by differences in reporting by access group, we validated self-report data using confirmed conditions in the EHR. We did not find higher rates of misreport among those with lower access to care, supporting our conclusion that the observed differences were related to EHR reliability rather than reporting. We were also unable to validate the accuracy of the outcome label used in the prediction task. As noted earlier, we selected T2D because it had the highest EHR reliability among the conditions considered; however, this label may still introduce noise, particularly if it is less precisely measured among participants with lower access to care, which may lead to an underestimation of differences in predictive performance among participants with lower access to care. EHR data were limited to clinical sites participating in the All of Us Research Program. Although this restricts observability of patient data across other healthcare settings, it mirrors the type of data typically available to health systems when developing clinical models, making it more representative of real-world prediction scenarios. The All of Us Research Program is

a convenient sample and not a representative sample of patients in the United States more broadly. While All of Us participants appeared to be from more diverse, lower income backgrounds compared to the ACS sample, restricting to participants that responded to the health access survey increased the percent participants that were white, had insurance, and higher income. This restriction may exclude some of the participants most at risk of being underrepresented in clinical prediction.

In this paper, we highlight the potential limitations of relying on EHR data for clinical risk prediction: patients with lower access to care have higher rates of missing data and are at risk of lower model performance. By quantifying this phenomenon at scale, we provide the first empirical evidence that healthcare access disparities directly undermine the promise of AI-driven precision medicine – highlighting that algorithmic fairness cannot be achieved without addressing fundamental access inequities.

Methods

This study was approved by the Tufts IRB and deemed Not Human Subjects research.

Data & Sample

Our data come from the All of Us Research Program.[12] The All of Us Research Program is a convenient sample of over a million people across 50 states. Data include survey data and (contingent on participant agreement) linked EHR data from participating provider organizations in the All of Us Research Program. Several surveys are administered to participants upon their enrollment in All of Us. Surveys include questions on participant demographics (“The Basics” survey), health history (“Personal and Family Health History” survey), and health care access (“Health Care Access & Utilization” survey). EHR data are standardized using the Observational Medical Outcomes Partnership (OMOP) Common Data Model infrastructure and are current through October 1, 2023. Our analysis sample includes all participants with linked EHR data that answered the “Health Access & Utilization” survey.

Participant responses from the “Health Access & Utilization” survey are used to define access groups. Participants responded to 9 questions related to delayed care and 15 questions related to affordability of care. We define participants as having delayed care if they responded that they had delayed care for any of the 9 reasons listed. We define participants as having cost-constrained care if they responded affordability concerns for any of the 15 areas. Participants that indicated they were unable to afford care on three or more questions were defined as having “severe cost-constrained care”. Similarly, patients that indicated they had delayed care on three or more questions were defined as having “severe delayed care”. We define participant age, gender, race, and ethnicity using data collected from “The Basics” survey. Participant age was calculated using the date of birth reported by the participant. Participant gender was measured based on the participants’ description of their gender identity. We defined a participant as Black if they responded “Black, African American, or African” to the question “Which categories describe you?”. We defined a participant as Hispanic if they responded “Hispanic, Latino, or Spanish” to the same question. Participants could indicate multiple categories.

Self-reported condition data was reported in the All of Us “Personal and Family Health History” survey. For each condition, the participant was asked the following question: “Including yourself, who in your family has had [condition name]”, where the name of the condition was provided within the brackets? Respondents had the choice to answer Self, Mother, Father, Sibling, Daughter, Son, Grandparent, or skip the question. We defined a participant as having a condition if they responded “Self” to the question. In total, we measured responses to 37 self-reported conditions (listed in Supplementary Table 1). Age-adjusted prevalence rates of conditions were constructed using 10-year age buckets.

Comparison to the American Community Survey

To understand how the All of Us participants and more specifically our analytic sample differed from the U.S. population more broadly, we compared sample characteristics to 1-year 2023 estimates reported by the U.S. American Community Survey (ACS). In order to make an accurate comparison, we used the same definition as the ACS for defining age categories (18–24, 25–44, 45–64, 65+), gender (male, female), race and ethnicity (AI/AN, Asian, Black, Hispanic or Latino, Other, White), health insurance (employee, Medicaid only, Medicare, Other insurance including purchased insurance plans, Tricare, or Veterans Affairs), and annual household income (<$10K, $10,000-$49,999, $50,000-$99,999, $100,000-$199,999, >=$200,000). All of Us participants with unknown responses within each category were removed for the purposes of comparison.

Comparing Self-Reported versus EHR-Reported Conditions

To test rates of concordance by access to care, we compared self-reported conditions to health conditions documented in the EHR. We identified EHR conditions by searching for all condition concepts (defined using the OMOP format) that matched the given condition reported in the survey. In Supplementary Table 1, we provide a list of concept codes used to identify conditions in the EHR data and the relevant survey question used to identify conditions in the survey data. We quantified the percent of participants with a self-reported condition that had no record of the same condition in the EHR (or “missing EHR diagnosis rate") for 37 self-reported conditions (Supplementary Table 1). The EHR reliability rate is the complement of this, or one minus the missing EHR diagnosis rate. Reliability rates for the standard care group were compared to reliability rates for the cost-constrained and delayed care group, respectively, using two-proportions z-test comparisons. We additional quantified the percent of participants with an EHR-reported condition (as of the time the self-reported condition data was collected) that did not self-report the health condition, among those with standard care versus non-standard care. Comparisons between the two groups were compared using a two-proportions z-test. P Values for comparisons were adjusted using the Benjamini-Hochberg procedure and statistical significance defined at the adjusted P Value level of 0.05.

To assess how standard care modifies the relationship between self-reported conditions and EHR-reported conditions, we fit an ordinary least squares regression model predicting the presence of a condition in the EHR among participants that self-reported any of the 37 conditions. For each participant i, we included a vector of 37 indicator variables

$$\boldsymbol{D_i} = (d_{i1},\ d_{i2},\ \ldots,\ d_{i37}),$$

Where $d_{ij} = 1$ if the participant reported condition j and $d_{ij} = 0$ otherwise. These indicators are included to account for variation in reliability among different types of condition. We also included an indicator for having standard care, $S_i = 1$ if participant i is in the standard care group and $S_i = 0$ if they are not. The final model specification is:

$$Y_i = \alpha_0 + \alpha_S \mathrm{S_i} + \boldsymbol{\alpha}^T \boldsymbol{D_i} + \epsilon,$$

where $Y_I = 1$ indicates that the condition appeared in the EHR and $Y_I = 0$ indicates that it did not. The coefficient of interest is the coefficient on the indicator for standard care α_S, which represents the percentage change in the probability that a condition appears in the EHR for someone with standard care compared to someone without standard care, controlling for condition type.

In a second specification, we included the number of visits a participant had recorded in the EHR data V_i, to control for visit intensity:

$$Y_i = \alpha'_0 + \alpha'_S \mathrm{S_i} + \boldsymbol{\alpha'}^T \boldsymbol{D_i} + \gamma V_i\ \ \epsilon,$$

and compared α_S to α'_S to assess how visit intensity modified the relationship between access to care and EHR reliability.

Implications for Clinical Prediction Models

The diabetes sample included any participant with no EHR-record of T2D as of the time in which the participant completed the “Personal and Family Health History” survey. We refer to this date as the index date. Using HbA1c lab values, fasting glucose lab values, random glucose lab values, and condition codes, we determined whether someone had diabetes. Lab values could take on a measure of 1 if met clinical criteria for diabetes, 0 if did not meet clinical criteria for diabetes, or NA if a lab record existed but the result was not recorded.

The outcome is an indicator for whether someone had a record of T2D in the two years following the index date (1 = Yes, 0 = No). A participant was considered to have T2D if they met any of the following criteria: at least two random glucose levels equal to or greater than 200 mg/dL, fasting glucose level equal to or greater than 126 mg/dL, HbA1c level greater than or equal to 6.5%, or a type 2 diabetes concept ID of 201826 (based on the OMOP data model).

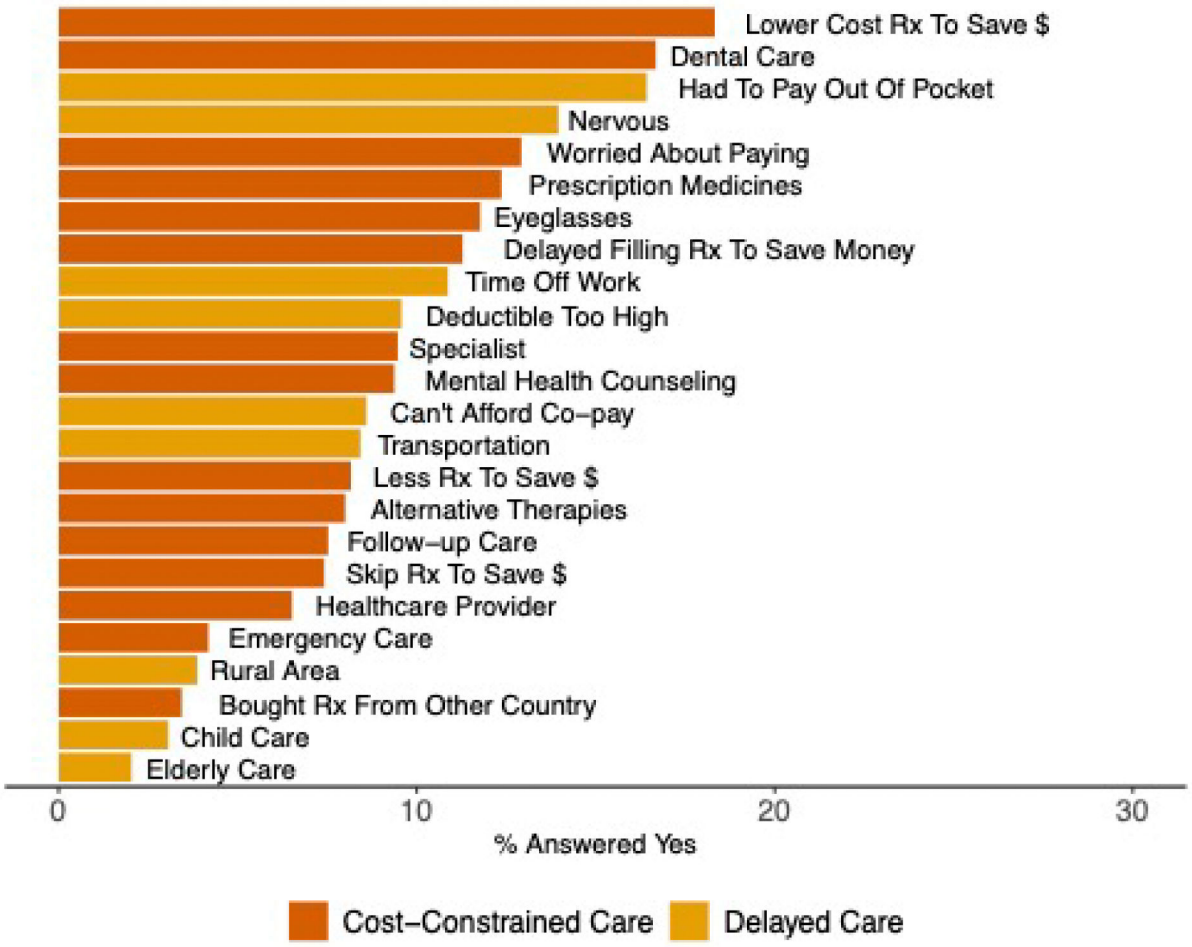
Features are constructed using a two-year lookback window from the index date. Baseline features included indicators for clinical conditions in the EHR data with prevalence rates of at least 1% in the lookback window, indicators for ATC-2 medications in the EHR data with prevalence rates of at least 1% in the lookback window, clinical measures in the EHR

(obesity, diabetes, fever, hypocalcemia, hypercalcemia, hypochloremia, hyperchloremia, creatine, high blood pressure, high triglyceride, tachycardia, anemia, high hemoglobin, hyperkalemia, hypokalemia, tachypneic, bradypnea, hypernatremia, hyponatremia, low urea, and high urea, see Supplementary Table 2 for lab values and ranges for each measure), and demographics (normalized age, female, male, white, black, Hispanic). All features except for age were constructed as indicator (Yes = 1, No = 0) variables. Missing data for indicator variables were set to zero. There was no missing data on age.

The baseline+access model included the same set of features as the baseline model plus two access variables: an indicator of a participant reported cost-constrained care and an indicator if a participant reported delayed care. The baseline+SES model included indicators for health insurance categories (employee, Medicaid only, Medicare, Other insurance including purchased insurance plans, Tricare, or Veterans Affairs), household income (<$10K, $10,000-$49,999, $50,000-$99,999, $100,000-$199,999, >=$200,000), zip-level socioeconomic status data from the 2017 ACS (the fraction of population with income in the past 12 months below the poverty level, the fraction of houses that are vacant, the fraction of the population 25 and older with an education attainment of at least high school graduation, and the deprivation index), and a flag if the patient hadn’t visit in the prior two years (during the window in which the predictor variables were constructed). Missing data on zip-level features was replaced using mean imputation. All continuous variables (age and zip-level features) were scaled and normalized. Missing data for indicator features was set to 0.

For the baseline prediction task, logistic regression and random forests were trained with 5-fold cross validation. Logistic regression models were developed using a regularization parameter search performed by the *cv.glmnet* package in R. Random forest models were developed with a regularization parameter search over 5, 50, 500, and 1000 decision trees, and the minimum node parameter set to 10. We use the model with the best AUC value for the final predictive task and all subsequent models. Model performance (AUC, balanced accuracy, sensitivity, and specificity) was assessed overall and by access levels. For classification, we selected the cutoff point that maximized Youden's J statistic. For uncertainty quantification of the classification metrics, we compute the numerical 95% confidence intervals as derived from the Central Limit Theorem. We use DeLong’s method to compare AUC values.[29]

Extended Data



Extended Data Figure 1. Share of Participants who Responded Yes to Access-Specific Questions

Responses from 305,857 participants that responded to the Health Access & Utilization survey. We define participants as having low access if they indicated they had cost-constrained or delayed care for any of the listed reasons.

Extended Data Table 1.

Sample Characteristics Compared to the U.S. Population

	Analytic Sample	All of Us Participants	U.S. ACS (1-year 2023 estimate)
Age (%)			
18–24	3.1	4.6	11.6
25–44	26.2	31.3	34.4
45–64	33.1	34.5	31.4
65+	37.6	29.6	22.6
Gender(%)			

	Analytic Sample	All of Us Participants	U.S. ACS (1-year 2023 estimate)
Male	35.1	37.0	49.5
Female	64.9	63.0	50.5
Race and Ethnicity (%)			
AI/AN	0.5	1.1	0.5
Asian	3.0	3.5	5.9
Black	9.9	15.7	11.8
Hispanic or Latino	13.5	18.1	19.4
Other	5.2	5.9	5.2
White	67.9	55.7	57.1
Health Insurance			
Employee	41,2	36.6	46.8
Medicaid Only	13.4	19.8	9.7
Medicare (Duals, supp.)	32.4	26.1	19.6
Other (Purchased, Tricare, VA)	8.9	9.9	15.2
Uninsured (inc. IHS)	4.1	7.6	8.7
Annual Household Income			
<$10K	7.7	14.4	5.3
$10,000–$49,999	28.9	32.9	23.7
$50,000–$99,999	28.4	24.8	32.1
$100,000–$199,999	24.4	19.6	26.5
>=$200,000	10.6	8.3	12.4

All of Us sample characteristic information is based on self-reported information found in the All of Us baseline survey. The Other Race category includes people who answered, ‘More than one population’,’None of these’,’Middle Eastern or North African’,’Native Hawaiian or Other Pacific Islander’. Unknown race and ethnicity categories include participants who skipped or selected that they preferred not to answer. AI/AN = American Indian or Alaska Native. Medicare insurance includes participants that are dually enrolled in Medicaid and those with supplemental insurance. Other insurance includes participants with directly purchased health insurance, TRICARE, Veterans Affairs (VA), and other military insurance. Those with Indian Health Service (IHS) were included in the uninsured category following the categorization used by the American Community Survey (ACS). All of Us participants with unknown responses within each category have been removed for purposes of comparing to the ACS.

Extended Data Table 2.

Age-Adjusted Self-Reported Condition Prevalence Rate

Self-Reported Condition	Standard Care (N=94,310)	Cost-Constrained or Delayed Care (N=87,921)
Acid Reflux	35.6	45.0
Astigmatism	36.8	40.0
Hypercholesterolemia	34.9	37.4
Hypertension	34.0	39.3
Depression	30.2	46.0
Anxiety and/or Panic Disorder	21.5	33.9
Cataracts	22.6	24.9
Osteoarthritis	20.9	26.1
Asthma	19.8	27.4
Anemia	18.9	27.1

We report age-adjusted prevalence rates for the top ten most common self-reported conditions. For age adjustment, we find the prevalence rate of each self-reported condition within 10-year age intervals (20, 30, 40, 50, 60, 70, 80, 90, 100+). We calculate the weights for each age group based on the sample age distribution. The final age-adjusted prevalence is the sum of each age interval multiplied by its relevant weight. We used a logistic regression to assess the statistical significance of the relationship between access and self-reported condition, adjusting for age. P-values were adjusted for multiple comparisons based on Benjamini-Hochberg procedure.

Extended Data Table 3.

Diabetes Sample Characteristics

	Overall (N=93,593)	Standard Care (N=42,993)	Cost-Constrained Care (N=40,418)	Delayed Care (N=33,495)
Access Measures (%)				
Cost-Constrained Care	43.2	0.0	100.0	69.6
Delayed Care	35.8	0.0	57.7	100.0
(Average) Age (Years)	55.5	60.4	51.8	47.8
Gender (%)				
Male	31.9	38.1	26.3	24.2
Female	66.7	60.9	71.9	73.6
Other	1.4	1.0	1.9	2.3
Race (%)				
AI/AN	0.5	0.3	0.6	0.8
Asian	3.4	3.2	3.3	3.7
Black	8.2	6.8	9.7	9.9
White	72.8	78.4	67.1	66.4
Other	5.3	3.9	6.7	7.3
Unknown	8.7	5.9	11.7	11.2
Ethnicity (%)				
Hispanic or Latino	11.0	7.5	14.7	14.7
Not Hispanic or Latino	86.7	90.1	83.0	83.3
Unknown	2.3	2.3	2.3	2.1
Health Insurance				
Employee	47.1	49.9	42.0	47.9
Medicaid Only	11.3	6.6	16.3	16.6
Medicare (Duals, supp.)	25.5	33.2	20.1	12.9
Other (Purchased, Tricare, VA)	15.0	16.5	14.2	12.3
Uninsured (inc. IHS)	3.2	1.3	5.6	5.5
Unknown	0.5	0.4	0.5	0.5
Annual Household Income				
<$10K	5.8	3.2	8.6	9.0
$10,000–$49,999	24.0	16.4	33.0	31.6
$50,000–$99,999	25.9	26.6	25.5	25.0
$100,000–$199,999	23.1	28.2	16.9	18.0
>=$200,000	10.5	10.6	11.0	10.1
Unknown	2.9	2.7	3.3	2.9
Outcome (%)				
2-Year Diabetes Incidence	2.9	2.7	3.3	3.0

Sample includes All of Us participants that answered the Healthcare and Access survey and had linked EHR data. Survey responses were used to classify participants into three access groups: participants with cost-constrained and/or delayed care indicated they had trouble affording care and/or had delayed care in the past 12 months (participants can be represented in both the cost-constrained and delayed care groups), and participants that indicated neither were classified as having standard care. All sample characteristic information is based on self-reported information found in the All of Us baseline survey. The Other Race category includes people who answered, 'More than one population','None of these','Middle Eastern or North African','Native Hawaiian or Other Pacific Islander'. Unknown race and ethnicity categories include participants who skipped or selected that they preferred not to answer. AI/AN = American Indian or Alaska Native. Medicare insurance includes participants that are dually enrolled in Medicaid and those with supplemental insurance. Other insurance includes participants with directly purchased health insurance, TRICARE, Veterans Affairs (VA), and other military insurance. Those with Indian Health Service (IHS) were included in the uninsured category following the categorization used by the American Community Survey.

Extended Data Table 4.

Comparing AUC Values for Standard Care versus Other Access Groups

Access Groups	AUC	P-Value for Difference
Standard Care	0.763	-
Cost-Constrained Care	0.754	0.354
Delayed Care	0.768	0.596
Severe Cost-Constrained Care	0.752	0.603
Severe Delayed Care	0.770	0.393

This table reports AUC values from the baseline model. We tested for statistically significant differences in the AUC value in the standard care group versus the other access groups using DeLong's method in the *pROC* package in R.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We gratefully acknowledge *All of Us* participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's *All of Us* Research Program for making available the participant data examined in this study. This study used data from the *All of Us* Research Program's Controlled Tier Dataset v8, available to authorized users on the Researcher Workbench. IC, HZ, and AZ were funded by a Google Research Scholar award for this work. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. IC receives additional support from an Apple Machine Learning Faculty Research Award.

Data Availability

Due to Data Use Agreement (DUA) restrictions to protect health information, the data are available only to registered users of the All of Us Researcher Workbench.

Code Availability

Relevant code is hosted on the All of Us platform; we have also published the program files to a GitHub repository for public use: https://github.com/zinka88/access.

References

1. van Doorslaer E, Masseria C, Koolman X, & OECD Health Equity Research Group. Inequalities in access to medical care by income in developed countries. CMAJ 174, 177–183 (2006). [PubMed: 16415462]

2. Caraballo C et al. Trends in Racial and Ethnic Disparities in Barriers to Timely Medical Care Among Adults in the US, 1999 to 2018. JAMA Health Forum 3, e223856 (2022). [PubMed: 36306118]
3. Lavergne MR et al. Disparities in access to primary care are growing wider in Canada. Healthc Manage Forum 36, 272–279 (2023). [PubMed: 37340726]
4. Palm W et al. Gaps in coverage and access in the European Union. Health Policy 125, 341–350 (2021). [PubMed: 33431257]
5. Goldstein BA, Navar AM, Pencina MJ & Ioannidis JPA Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 24, 198–208 (2017). [PubMed: 27189013]
6. Davenport T & Kalakota R The Potential for Artificial Intelligence in Healthcare. Future Healthc J 6, 94–98 (2019).
7. Ghassemi M et al. A Review of Challenges and Opportunities in Machine Learning for Health. AMIA Jt Summits Transl Sci Proc 2020, 191–200 (2020). [PubMed: 32477638]
8. Shen JH, Raji ID & Chen IY The Data Addition Dilemma. Preprint at 10.48550/arXiv.2408.04154 (2024).
9. Sauer CM et al. Leveraging electronic health records for data science: common pitfalls and how to avoid them. Lancet Digit Health 4, e893–e898 (2022). [PubMed: 36154811]
10. Bower JK, Patel S, Rudy JE & Felix AS Addressing Bias in Electronic Health Record-Based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise. Curr Epidemiol Rep 4, 346–352 (2017). [PubMed: 31223556]
11. Beaulieu-Jones BK et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? npj Digit. Med. 4, 1–6 (2021). [PubMed: 33398041]
12. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY & Ghassemi M Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27, 2176–2182 (2021). [PubMed: 34893776]
13. Obermeyer Z, Powers B, Vogeli C & Mullainathan S Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453 (2019). [PubMed: 31649194]
14. Chen IY, Joshi S & Ghassemi M Treating health disparities with artificial intelligence. Nat Med 26, 16–17 (2020). [PubMed: 31932779]
15. Chen IY, Szolovits P & Ghassemi M Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA Journal of Ethics 21, 167–179 (2019).
16. Chen IY et al. Ethical Machine Learning in Healthcare. Annual Review of Biomedical Data Science 4, null (2021).
17. Murray J The 'All of Us' Research Program. N Engl J Med 381, 1884 (2019).
18. Razavian N et al. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data 3, 277–287 (2015). [PubMed: 27441408]
19. Park JH et al. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. npj Digit. Med. 3, 1–7 (2020). [PubMed: 31934645]
20. Zhang Y et al. Applying Artificial Intelligence Methods for the Estimation of Disease Incidence: The Utility of Language Models. Front. Digit. Health 2, (2020).
21. Delpino FM et al. Machine learning for predicting chronic diseases: a systematic review. Public Health 205, 14–25 (2022). [PubMed: 35219838]
22. Pratley RE The Early Treatment of Type 2 Diabetes. The American Journal of Medicine 126, S2–S9 (2013). [PubMed: 23953075]
23. Ong KL et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet 402, 203–234 (2023).
24. Lindström J & Tuomilehto J The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 26, 725–731 (2003). [PubMed: 12610029]
25. Lugner M, Rawshani A, Helleryd E & Eliasson B Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. Sci Rep 14, 2102 (2024). [PubMed: 38267466]

26. Mani S, Chen Y, Elasy T, Clayton W & Denny J Type 2 Diabetes Risk Forecasting from EMR Data using Machine Learning. AMIA Annu Symp Proc 2012, 606–615 (2012). [PubMed: 23304333]
27. Anderson JP et al. Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records. J Diabetes Sci Technol 10, 6–18 (2015). [PubMed: 26685993]
28. Lai H, Huang H, Keshavjee K, Guergachi A & Gao X Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders 19, 101 (2019). [PubMed: 31615566]
29. DeLong ER, DeLong DM & Clarke-Pearson DL Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44, 837–845 (1988). [PubMed: 3203132]
30. Zink A, Obermeyer Z & Pierson E Race adjustments in clinical algorithms can help correct for racial disparities in data quality. Proc Natl Acad Sci U S A 121, e2402267121 (2024). [PubMed: 39136986]
31. Call KT et al. Barriers to care in an ethnically diverse publicly insured population: is health care reform enough? Med Care 52, 720–727 (2014). [PubMed: 25023917]
32. Taber JM, Leyva B & Persoskie A Why do People Avoid Medical Care? A Qualitative Study Using National Data. J Gen Intern Med 30, 290–297 (2015). [PubMed: 25387439]
33. Nong P, Adler-Milstein J, Apathy NC, Holmgren AJ & Everson J Current Use And Evaluation Of Artificial Intelligence And Predictive Models In US Hospitals. Health Affairs 44, 90–98 (2025). [PubMed: 39761454]
34. Americans Give US Health Care System a C Grade, Calling for Policy Change. AJMC https://www.ajmc.com/view/americans-give-us-health-care-system-a-c-grade-calling-for-policy-change (2025).
35. Carter J, Skopec L, Buettgens M & Banthin J Uninsurance and Medicaid Eligibility among Young Adults in 2025: Patterns by State and Subgroup. https://www.urban.org/research/publication/uninsurance-and-medicaid-eligibility-among-young-adults-2025 (2025).
36. Rivera SC et al. Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. The Lancet Digital Health 5, e168–e173 (2023). [PubMed: 36828609]
37. Paudel R, Dias S, Wade CG, Cronin C & Hassett MJ Use of Patient-Reported Outcomes in Risk Prediction Model Development to Support Cancer Care Delivery: A Scoping Review. JCO Clin Cancer Inform e2400145 (2024) doi:10.1200/CCI-24-00145. [PubMed: 39486014]
38. McWilliams JM, Weinreb G, Landrum MB & Chernew ME Use Of Patient Health Survey Data For Risk Adjustment To Limit Distortionary Coding Incentives In Medicare. Health Affairs 44, 48–57 (2025). [PubMed: 39761451]
39. Getzen E, Ungar L, Mowery D, Jiang X & Long Q Mining for equitable health: Assessing the impact of missing data in electronic health records. Journal of Biomedical Informatics 139, 104269 (2023). [PubMed: 36621750]
40. Zhang H, Clark AS & Hubbard RA A Quantitative Bias Analysis Approach to Informative Presence Bias in Electronic Health Records. Epidemiology 35, 349–358 (2024). [PubMed: 38630509]
41. Wang HE et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc 29, 1323–1333 (2022). [PubMed: 35579328]
42. Dankwa-Mullan I et al. A Proposed Framework on Integrating Health Equity and Racial Justice into the Artificial Intelligence Development Lifecycle. Journal of Health Care for the Poor and Underserved 32, 300–317 (2021).
43. Wagaw F Linking Data From Health Surveys and Electronic Health Records: A Demonstration Project in Two Chicago Health Center Clinics. Prev. Chronic Dis. 15, (2018).
44. O'Brien EC et al. Concordance Between Patient-Reported Health Data and Electronic Health Data in the ADAPTABLE Trial. JAMA Cardiol 7, 1235–1243 (2022). [PubMed: 36322059]
45. Huang J, Galal G, Etemadi M & Vaidyanathan M Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. JMIR Med Inform 10, e36388 (2022). [PubMed: 35639450]

46. Martin LM, Leff M, Calonge N, Garrett C & Nelson DE Validation of self-reported chronic conditions and health services in a managed care population. Am J Prev Med 18, 215–218 (2000). [PubMed: 10722987]

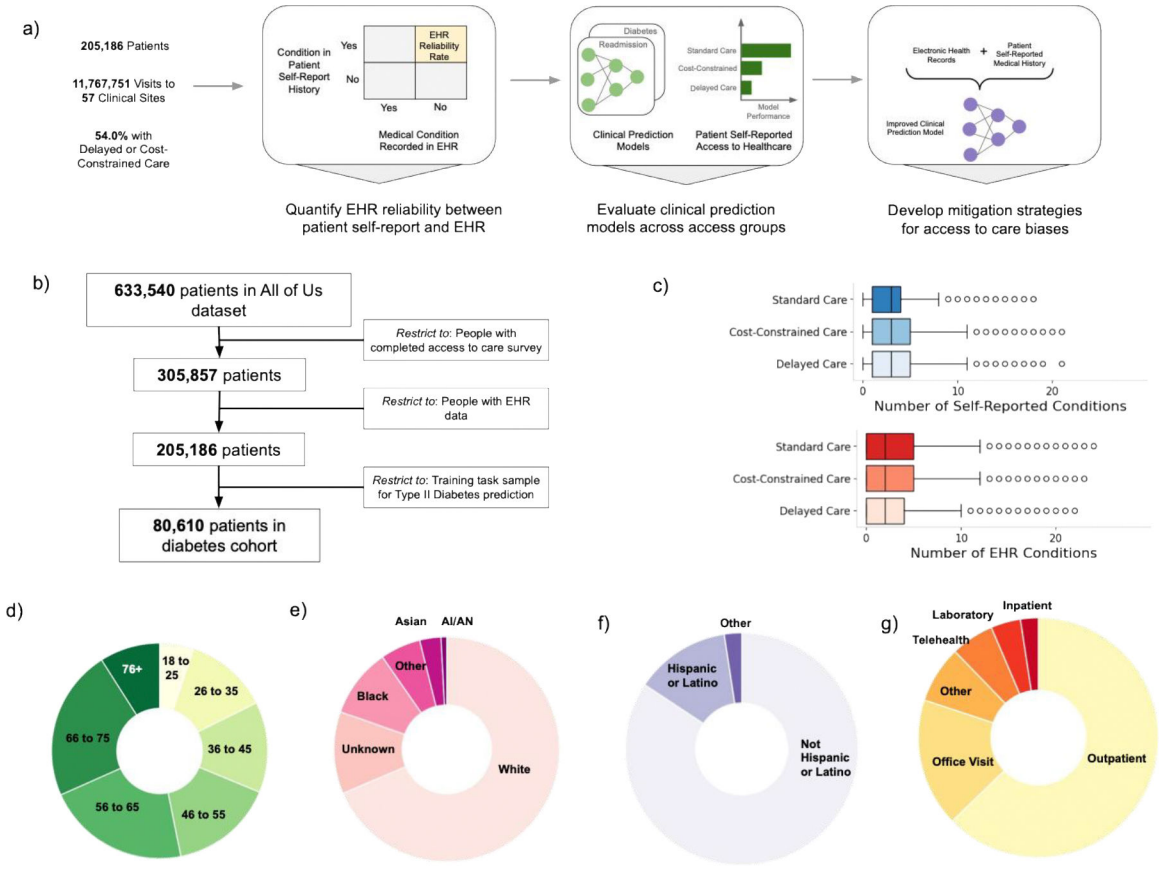


Figure 1:

Study overview and data cohort

Overview of paper approach to determine EHR reliability and clinical model performance bias due to differences in access to healthcare. (a) Using the All of Us dataset, we quantify EHR reliability by comparing patient self-reported medical history with electronic health record (EHR) medical conditions. We then assess the impact of differences in EHR reliability by evaluating trained clinical prediction models across access to care patient groups. Lastly, we develop mitigation strategies for these biases due to access to care using both EHR data and patient self-reported medical history. (b) Patient cohort creation flowchart beginning with the full All of Us patient dataset and excluding patients according to access to care survey completion and EHR data availability. (c) Distribution of self-reported diagnoses and EHR conditions per individual across access to care patient groups in our patient cohort. (d-f) Number of patients in each (d) age category, (e) race category, and (f) ethnicity category. (g) Number of EHR records in each visit type category.

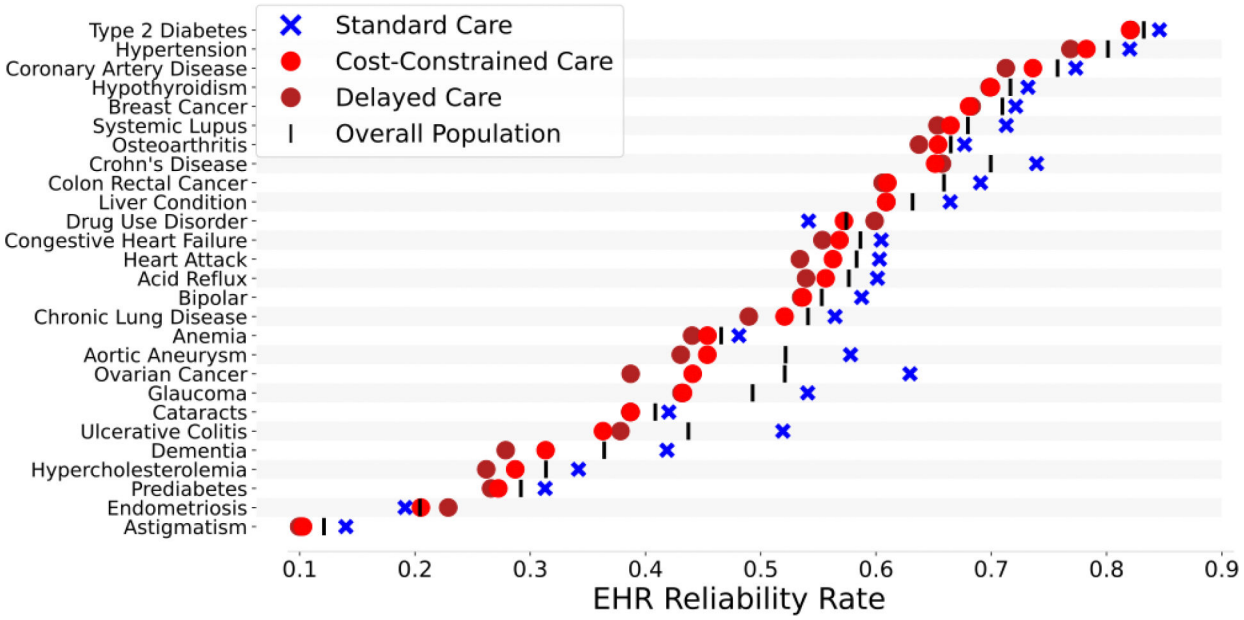


Figure 2:

EHR Reliability Rate for Conditions

Analysis of 205,186 participants (94,310 participants with standard care, 87,921 participants with cost-constrained care, and 73,445 participants with delayed care). EHR reliability was quantified by comparing patient self-reported medical history with electronic health record (EHR) medical conditions. We define the EHR reliability rate as the share of participants with a self-reported condition who had the condition documented in the EHR. In the figure we show all conditions with a statistically significant difference in missing EHR diagnosis rates between the standard care group and the cost-constrained or delayed care group. Statistical significance was calculated with a two-sided two-proportions z-test, with Benjamini-Hochberg adjusted P-values to account for comparisons for each condition within access group.

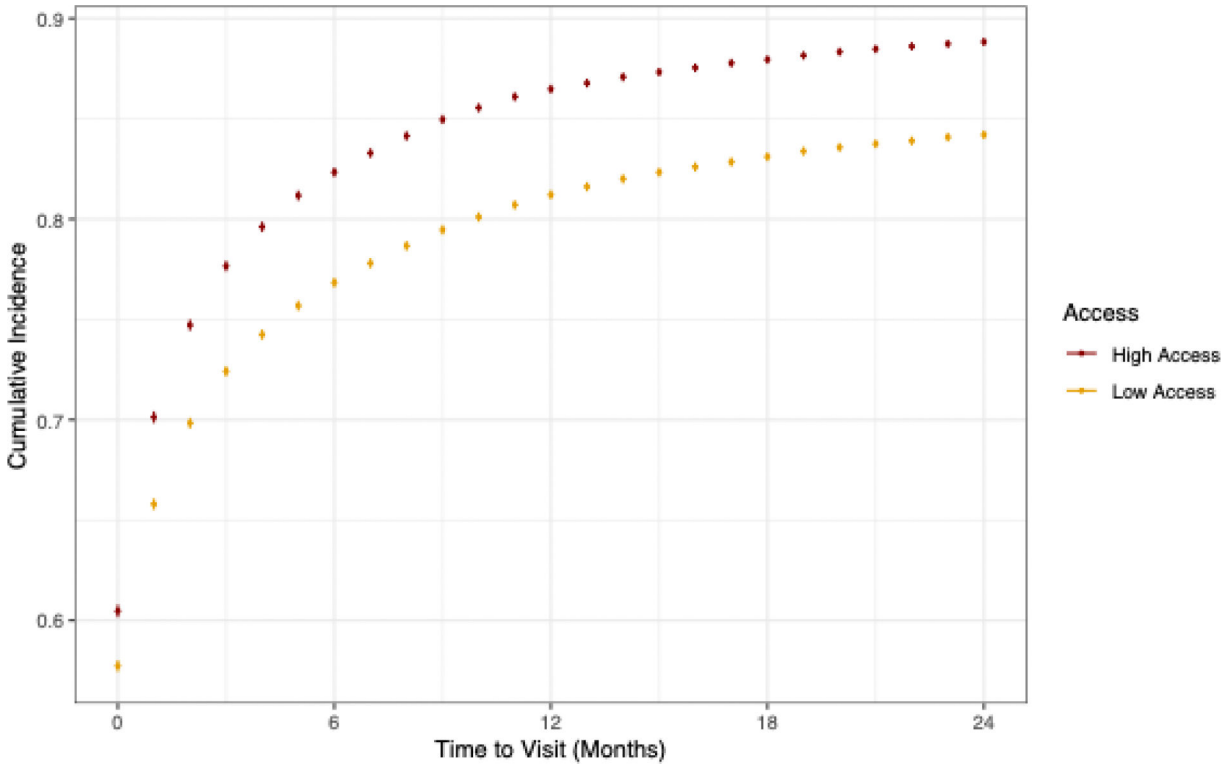


Figure 3:

Cumulative Incidence of Time to First Visit

Each point represents the cumulative percentage of participants within the specified access group who had a first visit documented in the EHR as of that point in time. Time was indexed such that time = 0 corresponds to the data in which sample participants answered the access survey. Standard errors for the cumulative percentages are calculated using Greenwood's formula and are shown as +/-1 standard error. The High Access group includes those with standard care (N=94,310). The Low Access group includes those with cost-constrained and/or delayed care (N=110,876).

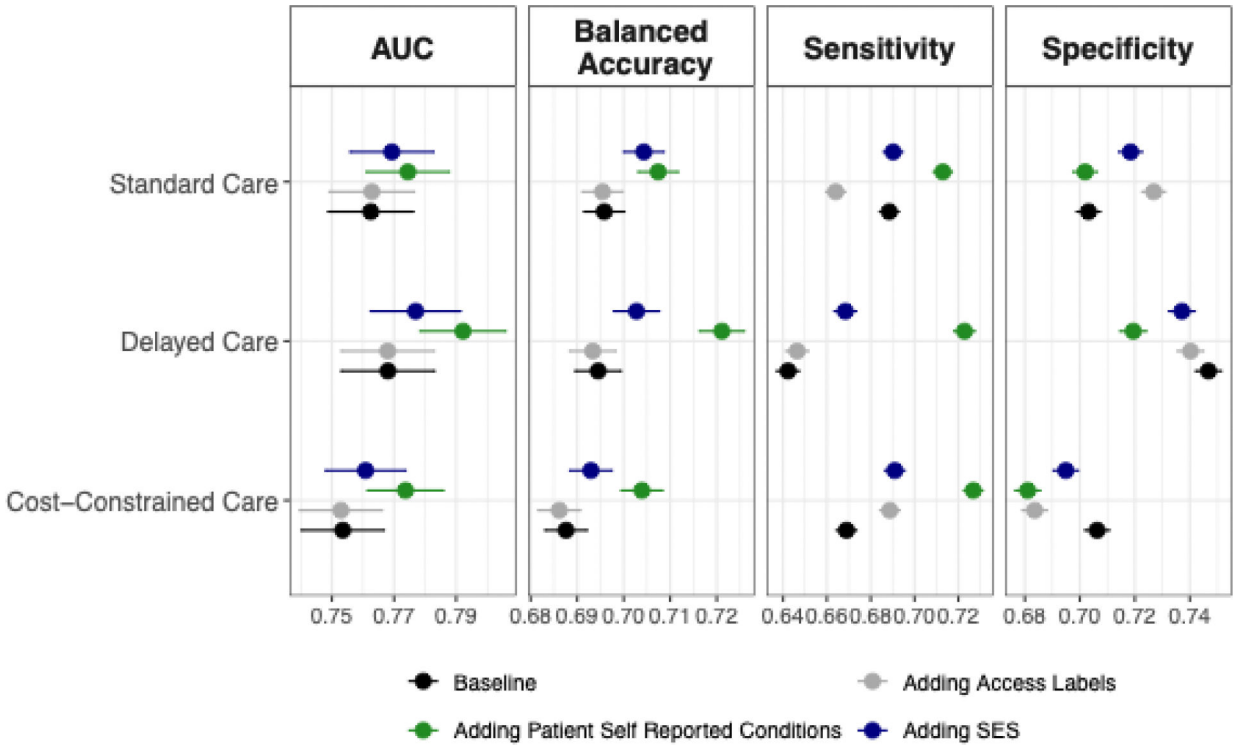


Figure 4:

Diabetes Task Model Performance

Analysis of 93,593 participants in the diabetes sample (42,993 participants with standard care, 40,418 participants with cost-constrained care, and 33,495 participants with delayed care). Each point represents a model performance measure. Error bars represent 95% confidence intervals. For AUC estimates, confidence intervals were derived using DeLong's methods.[29] For Balanced Accuracy, Sensitivity, and Specificity, confidence intervals were derived using normal approximation based on the Central Limit Theorem.

Table 1.

Study Sample Characteristics

	Overall (N=205,186)	Standard Care (N=94,310)	Cost-Constrained Care (N=87,921)	Delayed Care (N=73,445)
Access Measures (%)				
Cost-Constrained Care	42.8	0	100.0	69.1
Delayed Care	35.8	0	55.8	100.0
(Average) Age (Years)	55.7	60.4	52.1	47.8
Gender (%)				
Male	34.5	40.9	28.5	26.7
Female	63.8	58.0	69.1	70.6
Other	1.8	1.1	2.4	2.8
Race (%)				
AI/AN	0.8	0.6	1.0	1.2
Asian	3.0	3.0	2.9	3.3
Black	9.0	8.6	11.4	11.5
White	68.6	74.3	62.8	62.6
Other	5.8	4.3	7.3	7.8
Unknown	10.7	7.8	13.8	12.9
Ethnicity (%)				
Hispanic or Latino	13.3	9.5	17.1	16.7
Not Hispanic or Latino	84.3	88.0	80.4	81.1
Unknown	2.4	2.5	2.5	2.2
Health Insurance				
Employee	44.3	47.3	39.0	44.8
Medicaid Only	15.1	9.5	20.6	21.4
Medicare (Duals, supp.)	29.5	37.3	23.7	16.9
Other (Purchased, Tricare, VA)	17.3	19.6	15.3	14.0
Uninsured (inc. IHS)	3.7	1.7	6.2	5.9
Unknown	0.6	0.6	0.6	0.6
Annual Household Income				
<$10K	6.7	3.9	9.4	9.9
$10,000–$49,999	25.1	17.8	33.7	32.0
$50,000–$99,999	24.7	25.6	24.0	23.5
$100,000–$199,999	21.2	26.4	15.0	16.4
>=$200,000	9.2	13.3	4.3	5.7
Unknown	13.2	13.0	13.6	12.6

Sample includes All of Us participants that answered the Healthcare and Access survey and had linked EHR data. Survey responses were used to classify participants into three access groups: participants with cost-constrained and/or delayed care indicated they had trouble affording care and/or had delayed care in the past 12 months (participants can be represented in both the cost-constrained and delayed care groups), and participants that indicated neither were classified as having standard care. All sample characteristic information is based on self-reported information found in the All of Us baseline survey. The Other Race category includes people who answered, 'More than one population','None of these','Middle Eastern or North African','Native Hawaiian or Other Pacific Islander'. Unknown race and ethnicity categories include participants who skipped or selected that they preferred not to answer. AI/AN = American Indian or Alaska Native. Medicare insurance includes participants that are dually enrolled in Medicaid and those with supplemental insurance. Other insurance includes participants with directly purchased health

insurance, TRICARE, Veterans Affairs (VA), and other military insurance. Those with Indian Health Service (IHS) were included in the uninsured category following the categorization used by the American Community Survey.

Table 2.

Comparing AUC Values for Baseline vs Alternative Methods

	Baseline	Adding Access Features	Adding SES Features	Adding Self-Report Features
All	0.762	0.763 *	0.770 ***	0.780 ***
Standard Care	0.763	0.763	0.769 ***	0.774 ***
Cost-Constrained Care	0.754	0.753 *	0.761 ***	0.774 ***
Delayed Care	0.768	0.768	0.777 ***	0.792 ***

Analysis of 93,593 participants in the diabetes sample (42,993 participants with standard care, 40,418 participants with cost-constrained care, and 33,495 participants with delayed care). AUC comparisons between methods were calculated using two-sided DeLong's test for paired ROC curves in the *pROC* package in R. P Values were adjusted using a Benjamini-Hochberg procedures.

*** P value <0.001

** P value <0.01

* P Value< 0.05; . P Value< 0.10.