Bias Delayed is Bias Denied? Assessing the Effect of Reporting Delays on Disparity Assessments

Jennah Gosciak*
Cornell University
Ithaca, New York, USA
jrg377@cornell.edu

Aparna Balagopalan*
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
aparnab@mit.edu

Derek Ouyang
Stanford University
Stanford, California, USA
douyang1@stanford.edu

Allison Koenecke
Cornell University
Ithaca, New York, USA
koenecke@cornell.edu

Marzyeh Ghassemi
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
mghassem@mit.edu

Daniel E. Ho
Stanford University
Stanford, California, USA
dho@law.stanford.edu

Abstract

Prior work has documented widespread racial and ethnic inequities across sectors, such as healthcare, finance, and technology. Across all of these domains, conducting disparity assessments at regular time intervals is critical for surfacing potential biases in decision-making and improving outcomes across demographic groups. Because disparity assessments fundamentally depend on the availability of demographic information, their efficacy is limited by the availability and consistency of available demographic identifiers. While prior work has considered the impact of *missing* data on fairness, little attention has been paid to the role of *delayed* demographic data. Delayed data, while eventually observed, might be missing at the critical point of monitoring and action – and delays may be unequally distributed across groups in ways that distort disparity assessments. We characterize such impacts in healthcare, using electronic health records of over 5M patients across primary care practices in all 50 states. Our contributions are threefold. First, we document the high rate of race and ethnicity reporting delays in a healthcare setting and demonstrate widespread variation in rates at which demographics are reported across different groups. Second, through a set of retrospective analyses using real data, we find that such delays impact disparity assessments and hence conclusions made across a range of consequential healthcare outcomes, particularly at more granular levels of state-level and practice-level assessments. Third, we find limited ability of conventional methods that impute missing race in mitigating the effects of reporting delays on the accuracy of timely disparity assessments. Our insights and methods generalize to many domains of algorithmic fairness where delays in the availability of sensitive information may confound audits, thus deserving closer attention within a pipeline-aware machine learning framework.

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution 4.0 International License.

FAccT '25, Athens, Greece
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732123

CCS Concepts

• **Applied computing**; • **Computing methodologies** → **Modeling and simulation**; • **Social and professional topics** → **Race and ethnicity**;

Keywords

Healthcare, disparity assessments, audits, delayed reporting, missingness

ACM Reference Format:
Jennah Gosciak, Aparna Balagopalan, Derek Ouyang, Allison Koenecke, Marzyeh Ghassemi, and Daniel E. Ho. 2025. Bias Delayed is Bias Denied? Assessing the Effect of Reporting Delays on Disparity Assessments. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3715275.3732123

1 Introduction

Racial inequity in the United States (U.S.) remains a significant issue in sectors such as healthcare, employment, finance, and education.[1] In healthcare, where such disparities can be stark [5, 45, 52], researchers, policymakers, and healthcare institutions have increasingly turned toward assessments to measure, and potentially mitigate, such disparities [74]. Such assessments are also crucial tools for auditing the fairness of machine learning (ML)-based diagnostic tools — an area of growing concern as ML and data-driven decision-making become more prominent in healthcare [13].

Less recognized is a core impediment for disparity assessments: the *timely reporting of demographic information* (*e.g.*, race, gender) by patients and providers. In this work, we show that failing to account for reporting delays, as distinct from missing data, can obfuscate health disparities. Leveraging access to a large, longitudinal dataset of over 5M patients, sourced from primary care practices throughout the U.S., we both document the extent of race reporting delays and examine the effect on disparity assessments, which we expect to be increasingly common.

We make several contributions in this work. First, we provide researchers and practitioners with a concrete definition of *delay*,

[1] For brevity, we use "race" to refer to both race and ethnicity throughout the remainder of this paper.

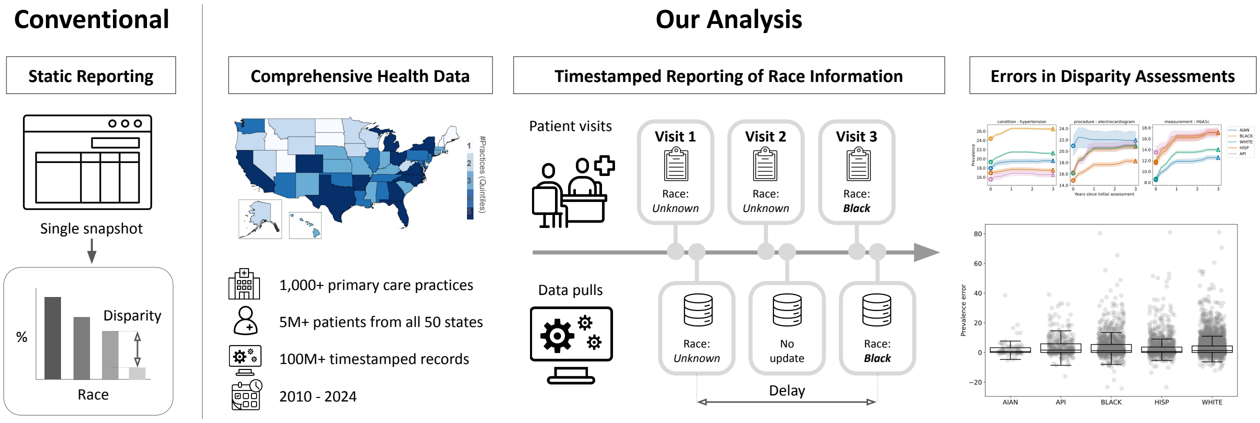

Figure 1: On the left, we show the *conventional* approach to disparity assessments, which results in a *static* measure of disparity. On the right, we present the three core components of our analysis. First, we leverage access to a *comprehensive* dataset of over 1,000 primary care practices, 5M patients from all 50 states, and 100M patient interactions from 2010 to 2024. Second, we use *timestamped* records to identify and measure delays in reporting of race information. Third, we demonstrate how reporting delays drive *errors* in disparity assessments across a variety of consequential health outcomes, from the national (Figure 3) to the practice level (Figure 11).

which occurs when information is initially unreported for an individual, but eventually becomes available after repeated interactions with data collection systems (*e.g.*, repeated patient visits to a primary care provider). Importantly, such delays may affect variables that are considered "static" [18] (*e.g.*, data usually collected at the time of hospital admission such as race and pre-existing diagnoses [63]). No prior works in healthcare or fair machine learning, to the best of our knowledge, have rigorously analyzed the impact of this type of temporal missingness of demographic attributes in administrative data, which we are able to observe through a richly timestamped healthcare dataset. We show that, in fact, delays are widespread. Second, we examine heterogeneity in reporting delays and find that rates of delayed reporting vary by race (and other healthcare attributes), directly implicating bias concerns. Third, we design and carry out a series of retrospective analyses on this data to understand how delayed race reporting impacts disparity assessments in a real-world, high-impact setting. We find consequential distortions, with prevalence errors of 10 percentage points or more not uncommon at the practice level. Lastly, we demonstrate that widely used imputation methods like Bayesian Improved First Name Surname Geocoding (BIFSG) [84], while relatively accurate at individual prediction of race, do not significantly reduce errors in disparity assessment across all outcomes of interest.

Our work highlights the importance of pipeline-aware, context-specific approaches to data-driven decision making [4, 16, 81]. Pipeline-aware fairness involves considering all of the different design decisions in the full ML pipeline and their effect on fairness outcomes. As Black et al. [16] demonstrate, far more effort has been spent studying bias in statistical models. Much less attention has been paid to other aspects of the ML pipeline such as data collection – the focus of our paper. In settings involving time-sensitive, routine disparity assessments (*e.g.*, dashboards measuring health outcomes for different racial groups), delayed race data may hinder *responsive* and *actionable* feedback. Our results suggest researchers and practitioners should expend greater efforts to identify sources of delay that might exist within real-world data collection pipelines, consider their downstream impacts, and test policy and/or programmatic interventions to reduce delays. As we show, delayed reporting may lead to inaccurate and misleading estimates of disparities, with direct fairness implications; these findings may similarly affect other high-stakes applications and geographic domains. Figure 1 summarizes the value of our analytic approach, which surfaces errors in disparity assessments across time and geography by leveraging timestamped race reporting information from a unique health dataset, all of which would not be possible via a conventional approach to disparity assessments using more static data.

The rest of the paper is structured as follows. In Section 2, we discuss the changing policy landscape related to monitoring and addressing health disparities in the U.S. We also connect our work to practical challenges in algorithmic fairness in the wild: *(1)* the frequency of missing demographic data in many real-world contexts, and *(2)* the importance of studying fairness *dynamically*. Section 3 provides an overview of our dataset, which uniquely affords us access to information from over 1,000 practices across all 50 states and over 5M patients in the U.S. Most notably, the data contains fine-grained longitudinal information across over 100M patient interactions, including timestamped reporting of race information, which enables us to design realistic assessments of the magnitude, correlates, and impact of reporting delays. Then, in Section 4 we detail our methods, including details on data processing, key definitions, and summary metrics. Section 5 describes the population of patients with and without delays and presents the results from our retrospective analyses conducted on real patient data. Lastly, in Section 6, we discuss the implications of our findings for researchers and practitioners in both healthcare and algorithmic fairness. In particular, we call attention to the importance of considering fairness in real-world deployment settings where the reporting mechanism for demographic attributes may lead to delays over time.

2 Background and Related Work

Prior research has extensively documented racial health disparities in the U.S., from pain management to life expectancy [6, 32, 41, 44, 66, 72]. However, there are several impediments to *accurate and timely assessments* of disparities. In our work, we focus on one challenge that has been neglected in prior work, but is of immense practical consequence: reporting delays in demographic information. While several prior works have studied system fairness over time, these often focus on the distribution shift [59] arising from changing sub-populations or systems behavior [70], whereas we focus on a setting where the population remains the same but data completion rates change over time. In the sections below, we describe both the policy background and the algorithmic fairness literature motivating this work.

2.1 Policy Background

Data Infrastructure for Measuring Health Care Disparities. Landmark studies, such as the "Heckler Report" (1985) [46] and the Institute of Medicine's (IOM) "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care" report [54], have centrally shaped our understanding of racial disparities in the U.S. healthcare system. Published nearly 20 years apart, these reports revealed the harmful impact of racial health disparities throughout the U.S., recommended improved data collection related to race, and advanced legal, regulatory, and policy interventions specific to the medical field. Following these reports, considerable research has focused on reducing health disparities [72]. In our work, we build on this research by advocating for the data infrastructure necessary for timely disparity assessments in primary care settings.

The current picture of health disparities — at the national, state, or local level — is limited by the quality of available demographic data [15]. Many challenges first identified by IOM persist, in part due to poor data collection [55]. A recent *Urban Institute* report from James et al. [55] details challenges to collecting data on race, including lack of trust from both patients and providers, limited community engagement, and fragmented and inconsistent data systems. Patients may be asked to provide their race multiple times, with varying standards for recording racial categories across institutions. Data collection efforts are frequently uncoordinated and siloed across different patient interaction points, such as hospitals and insurance plans. Often, collection of race information is simply not a priority, and there are few mechanisms in place to produce high quality demographic data collection. *Overall, no widespread requirements for standardization and timely reporting of race in healthcare organizations exist. In particular, the goal of conducting regular disparity assessments for meaningful health outcomes, while acknowledged as worthwhile, has not been a consistent policy priority* [27, 55].

Mandated Reporting at Federal and State Levels. Recent federally-led efforts to address such gaps have been slow and inconsistent [67]. For example, since 2016, Medicaid has required states to develop health disparity assessment plans, which would require stratification by race. However, implementation by the Centers for Medicare and Medicaid Services (CMS) was itself delayed, and reporting eventually became voluntary [85]. New regulations — 42 C.F.R. § 437.10(b)(7), (d) (2023) — will require states to report core health quality measures stratified by race starting in 2027 [85]. CMS also recently announced the *Hospital Commitment to Health Equity* measure, which requires that hospitals participating in CMS programs report on whether they are prioritizing equity, but does not require systematic health disparity assessments [26].

More concrete advances have been led by states. Since 2011, Michigan has been reporting on health disparities annually, with programmatic efforts to reduce these disparities. Michigan Medicaid now links reimbursements and performance bonuses to reductions in health disparities across five measures: diabetes (hemoglobin HbA1c) testing, cervical cancer screening, child wellness visits, postpartum care, and chlamydia screening, an approach that aligns with James et al. [55]'s recommendations. In particular, they argue for tying accountability measures and incentives to the reporting of health disparities. California now requires the reporting of race information, although data collection is fragmented and lacks a universal standard. Legislation enacted in 2022 requires hospitals to prepare and submit annual health equity reports along with an action plan [74]. The first reports will be due in mid-2025 and will involve annual reporting on health outcomes disaggregated by race, among other demographic characteristics [74]. As numerous other states and localities propose similar initiatives to regularly report on disparities [68], it will be increasingly important to consider, understand, and potentially mitigate the impact of data reporting delays on accurate assessments. Our study contributes a comprehensive framework towards these aims.

2.2 Machine Learning Background

Dynamic and Pipeline Aware Fairness. Our work is also related to dynamic or longitudinal fairness [28, 39] assessment where fairness of a sociotechnical system is assessed over time – due to shifting populations [59] or system updates [70]. However, in contrast to prior work, we highlight that *missingness* in critical data variables can occur dynamically due to delayed reporting (*e.g.*, of race). Another related literature is early stopping in clinical trials where preliminary measurements of health outcomes can lead to incorrect disparity assessments. Prior work [2, 24] has considered the fairness implications of early stopping and adaptivity in clinical trials. However, these settings differ from ours, in that it is assumed that demographic data is available for all patients upfront.

Our work underscores the importance of a *pipeline-aware* machine learning [16] perspective. In the context we study, a *pipeline-aware* perspective entails systematically interrogating data collection and reporting systems, as well as the way missing or damaged data is handled in data preprocessing, and how imputation might be performed. We identify delay as a potential blindspot in machine learning pipelines, as it is only recognizable across longitudinal snapshots of individuals, while the bulk of the literature has analyzed static datasets. Our work shows the need for explicitly examining reporting delays in decision-making pipelines.

Missingness in Demographic Information. Missing data and imputation are well-studied topics in statistics and sociology [64, 65]. The impact of missing data, particularly race data, has been widely studied in health research as well [15, 17, 29, 30, 38, 76, 79]. Within the algorithmic fairness literature, there has been considerable attention paid to the consequences of missing features – including sensitive information [3, 12, 57, 69, 88] and missing data imputation [36, 56, 89]. Zhang and Long [88] provide theoretical bounds on fairness estimation error in the presence of missing data. Fernando et al. [36] document several missingness patterns such as item non-response and attrition, and find that imputing rows with missing data can mitigate bias. However, they do not consider *delay* in their discussion of missing data.

Jeanselme et al. [56] study multiple forms of missingness processes and emphasize that no single imputation strategy outperforms across all processes. While they do not characterize delay, the implication of their findings is that delay, when it exists, would also manifest in its own unique patterns, further complicating efforts to address missingness through imputation. Akpinar et al. [3] study the systematic problem of "differential feature under-reporting": a phenomenon in which some data records are more likely to be complete for individuals who interact with the system more frequently. They show that under-reporting tends to exacerbate disparities and propose mitigation methods. Our work assesses whether delay is also differential in similar or dissimilar ways across patterns of care-seeking behavior.

Our focus specifically on delays provides a novel opportunity to advance the missing data literature. Not only are delays on their own an important source of missing data to consider in real-world applications, by definition, they produce data that is only missing for some period of time. In other words, the ability to validate the ground truth of delayed data might shed light on some of the mechanisms that contribute to missingness not at random (MNAR), which otherwise are not observable to researchers within a static dataset [12, 56, 57].

3 Data

We leverage access to the American Family Cohort (AFC) dataset, which contains data from over 1,000 practices, all of which are part of the American Board of Family Medicine (ABFM) PRIME Registry [80]. The PRIME registry functions as an intermediary between healthcare providers and the Centers for Medicare & Medicaid Services (CMS). They help with collecting and analyzing data, and produce quality measures on behalf of clinicians for incentive-based programs managed by CMS [25]. In contrast to many conventional machine learning datasets, these data contain fine-grained longitudinal information of patient interactions (see Figure 1), including changes in race reporting, which enables us to conduct realistic assessments of the magnitude, correlates, and impact of reporting delays.

Data Collection and Incentives for Disparity Assessments. Healthcare practices that join the PRIME registry have access to a detailed set of dashboards with information about their practice service area, disease prevalence, and care quality gaps. Many practices share data with the registry because they do not have the capacity to do their own analyses and reporting to be in compliance with CMS. As previously noted, while CMS does not yet require racial health disparities to be analyzed and reported, local programs and mandates are beginning to emerge. ABFM and partnering researchers are well-positioned to conduct disparity assessments using health data on behalf of practices in the registry. Our study is a practical demonstration of this, specifically the potential impacts of delayed reporting.

Key Features of Dataset. The AFC dataset is ideally suited for studying reporting delays. First, this data contains longitudinal information including patient demographics, visits, diagnoses, observations, and procedures, as well as some clinician-specific details. Second, practices from all 50 states are represented in the data. Third, the data includes significant representation from healthcare practices and patients with both private and public insurance plans, as well as distinct electronic health records (EHR) systems. These characteristics make AFC data a meaningful and realistic test case for studying racial disparities across the U.S., as opposed to analyses that may focus on less diverse sub-regions, specific providers, single EHR systems, or a small subset of medical conditions. Summary statistics for this dataset are in Table 5.

Most importantly for our study, the data is *longitudinal* and information updates are timestamped, providing the possibility of observing the phenomenon of delay that would otherwise be hidden. Every time information is modified or added for a patient, a new record is added to the AFC dataset with a timestamp and linkable patient ID (see Longitudinal Reporting in Figure 1), without overwriting previous timestamped records for the same patient. This includes cases when demographic data such as race is updated. As a result, we can track the reporting of race for each patient over time and produce estimates of delayed race reporting, differentiating this dataset from other datasets with a static availability of race per patient. These timestamps come from the data provider, and the cadence of the updates does not always follow a regular pattern. In particular, some practices *push* their data – meaning they submit the data to the registry – while others experience data *pulls* at regular intervals.

For large scale audits, by the time patient information is aggregated into the AFC dataset, *any upstream source of delay* in reporting of race — whether due to patient hesitance, failure to request the information at the time of the patient visit, or data collection lags on the part of the data provider — creates delay that can materially affect the quality of disparity assessments. Therefore, we focus on the *consequences* of delays rather than the precise *causes* of the delays. We further define reporting delays in the next section.

4 Methods

In this section, we describe how we define and quantify the impact of delays on disparity assessments[2].

4.1 Defining Delays

We define *race reporting delays* using a time-based measure of delay. Delayed reporting occurs if there is a gap between the earliest possible date of reporting, and when race is actually reported. We

[2] Code: https://github.com/reglab/delayed-reporting

operationalize the former as whichever occurs latest among (1) the earliest timestamp denoting when the patient's date of birth is reported or (2) the earliest per-practice race reporting (*i.e.*, the first patient in a practice with race information). We consider a patient's race to be *reported* if there is a non-missing race or ethnicity entry that corresponds to one of the federal race and ethnicity categories (as described in Section 4.2). Importantly, our date information reflects the date when this information was shared or updated with the data provider responsible for producing the AFC data, not the clinical encounter when the patient may have self-reported race. As a result, this date may lag in comparison to the true clinical encounter. However, this definition of delay still captures realistic data lags (*e.g.*, due to a range of behavioral, administrative, and technological factors) that an independent evaluator would encounter when conducting a disparity assessment.

4.2 Data Processing

4.2.1 Measurement of Race and Ethnicity. We parse and code categorical versions of patient race and ethnicity from both free-text and categorical race-related fields in the AFC data, following the same processing steps as Cheng et al. [22]. To harmonize across a wide range of data schemas, we map all entries to the 1997 Office of Management and Budget (OMB) federal standard for race and ethnicity reporting: American Indian or Alaska Native (AIAN), Asian, Black, Native Hawaiian or Pacific Islander (NHPI), White, Multiracial, Other, and Hispanic [75].[3] Following OMB standards, we record patients as Hispanic or Latino if they indicate their ethnicity as such, in addition to their indicated race. For analyses involving prevalence estimates, we combine Asian and NHPI into the Asian and Pacific Islander (API) category. To detect when race is unknown or declined, we use string matching to a curated set of keywords to identify data points with no reported race (see Appendix A). For patients who report race multiple times ($< 1\%$), we parse their first reported race — matching the time at which we consider race to be reported. Note that we expect detection rates for multi-racial patients to be lower than those for other racial groups, as only simple regex parsing rules are applied.

4.2.2 Cohort Definition. From the full AFC dataset of 7.8M patients, we restrict analysis to patients for whom we ever have a recorded race mappable to OMB categories, and we identify the earliest date at which that recorded race is available. Because our objective is to assess the prevalence and impact of *delays* (*i.e.*, cases for which we can eventually recover a race recording), we also exclude patients who *never* report race, or whose race cannot be parsed using our automatic processing techniques (~900k patients). We only consider patients who are ≥18 years in 2018, which is the primary year we use for most analyses. Our final cohort consists of $5,310,700$ adult patients whose race is recorded with either some or no delay (as of early 2024). We then identify whether a patient has experienced a reporting delay by producing a continuous measure reflecting the number of days from a patient's earliest reporting opportunity up until the date that race is in fact available (Section 4.1).

4.2.3 Health Outcomes and Metadata. We observe patient attributes that have been standardized and cleaned according to the Observational Medical Outcomes Partnership (OMOP) Common Data Model.[4] In addition to race (as described in Section 4.2.1), we extract patient age, sex, and marital status. We also extract clinical information such as the number of patient visits, the length of time they have interacted with a practice, and health-related outcomes like disease diagnoses, procedures, and observations. Like demographic characteristics, all of these features follow the OMOP data model. Like race information, some information is subject to reporting delays and missingness. For understanding the AFC population (see Table 5), we treat these attributes as fixed (*i.e.*, we extract attribute information if it ever appears in the AFC data) and do not consider the impact of delays beyond delays in race reporting.

In the context of disparity assessments, we compute six binary health outcomes: three condition diagnoses (depression, diabetes, and hypertension), two procedures (electrocardiograms and depression screens), and a clinical observation (hemoglobin HbA1c tests). We curate these health outcomes based on prior literature which provides evidence of racial disparities (see more details in Appendix B).

4.3 Retrospective Analysis on the Impact of Delays

Drawing on the definition of delay in Section 4.1 and the health outcomes described in Section 4.2.3, we next examine the impact of reporting delays via a retrospective analysis of disparities. We consider disparity assessments to include any comparison of health outcome prevalence by racial group, for a particular time period (*e.g.*, the White-Black hypertension diagnosis gap in the first quarter of 2018). Since we have timestamps of when race information became available for each individual patient, we can demonstrate what a particular disparity assessment would have looked like *retrospectively*, if it had been conducted at any previous time point. For instance, we can simulate a disparity assessment for 2018 Q1 immediately following its conclusion, at which point 40.94% of patients have delayed race information and are thus excluded from prevalence calculations by racial group. The same disparity assessment for 2018 Q1, conducted with the benefit of more hindsight (*i.e.*, using more complete race data provided after delay, but health outcomes remaining fixed for 2018 Q1), could yield different results because more patients would be included in the analysis given their race availability. Our core objective is to isolate this impact of reporting delays on health disparities.

We conduct these analyses at three distinct geographic levels: national, state, and practice-level. First, with simulations at the national level, our goal is to understand how delayed reporting of race may affect aggregated disparity estimates similar to annual reports like the "National Healthcare Quality and Disparities Report" [37]. Second, as noted above, specific states, such as California and Michigan, have been pushing for deeper assessments of health disparities, so we also conduct analyses at the state level. Simulating disparity assessments may yield more variation at the state level, particularly as some states may experience more delays than others. We may also more easily observe distinct trends in delayed

[3] Note that the race group of "Middle Eastern or North African" was only added in the 2024 OMB categories update [1].

[4] https://www.ohdsi.org/data-standardization/

reporting that are overridden at the national level. Lastly, because administration of race reporting and mitigation efforts occur within physician practices, we also study the impact of reporting delays at the practice level.

4.4 Metrics for Error in Disparity Assessments

In all cases, we measure prevalence, or rates of occurrence, of health outcomes. We introduce two primary error metrics of interest based on changing accuracy of health monitoring, as race information becomes more complete over time: *prevalence errors* and *disparity errors*. Prevalence errors are the differences in prevalence estimates for racial groups in a cohort (defined by a fixed time period such as 2018 Q1) at some initial time point t_{initial} (the first possible assessment of the cohort, when race information is most incomplete) compared to time point t_{final} (when all race information is known for the cohort).[5] We also visually present prevalence estimates at quarterly intervals past t_{initial} to show how error is reduced over time as more race information is collected (see Figures 3 and 4). Disparities are the pairwise comparisons of prevalences between two racial groups, and so disparity errors derive from, but do not necessarily appear the same as, prevalence errors (*i.e.*, prevalence errors in the same direction may yield no disparity error). We summarize all metrics in Table 1. To obtain uncertainty estimates, we bootstrap by resampling 50 times across both practices and individual patients, and averaging metrics across all bootstrapped samples.

5 Results

We begin by reporting results on the prevalence and correlates of delays. We then report results from our retrospective analysis described in Section 4.3 and calculate the error metrics described in Section 4.4. We find that the impact of reporting delays is substantial. Because of the richness of the dataset, we distill core results here, and provide more detailed results in the Appendix.

Delays are the norm, not the exception. Over 73% of patients ($N = 3,911,213$) in our cohort experience some delay, and over half experience delays ≥ 60 days. Overall, 21 states and over half of practices exhibit a similar degree of delay (75% of patients or more), indicating that the phenomenon is both widespread and consistent. Put differently, any well-intentioned efforts to conduct routine, quarterly assessments (*i.e.*, within three months of the health outcomes in question) would likely discard a majority of all patients from analysis.

Delayed reporting of race does not affect all groups evenly. If patients with timely reported data are representative of all patients, reporting delays may not pose a substantive problem. But, delays do not affect groups equally. Figure 2 shows that the cumulative rates of reporting are much steeper for racial groups like AIAN and NHPI, while White, Black, and Asian groups appear to experience greater lags. Kruskal-Wallis tests for all pairwise comparisons are statistically significant with $p < 0.01$, even with Benjamini-Hochberg adjustment for multiple comparisons. In short, reporting delays are not only pervasive, but themselves have a distributive dimension across race.

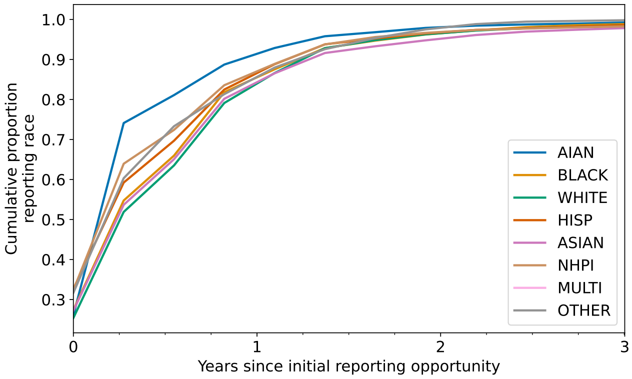

Figure 2: Reporting rates differ by race and ethnicity. On the x-axis, we show the time difference from the earliest date at which a patient could have reported race and ethnicity up until the date at which their race and ethnicity is *known*. To see the differences more clearly, we only depict the first three years on the x-axis. The cumulative proportion of patients within each racial and ethnic group who have reported race and ethnicity is on the y-axis.

Patients with delayed race reporting are older, more care-seeking, and less healthy. We also find that reporting delays are correlated with a wide range of other patient attributes, threatening the validity of static disparity assessments. Tables 2 and 5 present differences in means across a variety of patient-level characteristics, almost all of which are statistically significant. White patients are over-represented among those with delays while Hispanic patients are under-represented. Patients with delays also have more visits on average, and their first visit occurs earlier in the data. Lastly, patients with delays tend to be less healthy, with higher rates of all six health outcomes we measure. For example, they are more than 10 percentage points more likely to receive a diabetes (HbA1c) test.

Delay is also associated with differences in practice-level characteristics related to data collection and data management. Patients without delays are more likely to come from practices using Practice Management (PM) systems. PM systems automate many billing and administrative tasks, which can include data collection of demographic information [11]. This finding aligns with prior work that suggests integrating EHR and PM systems may lead to improved data collection [20]. Delays are also associated with practices that have their data only "pulled" — meaning the registry initiates extracting data on some regular cadence. Patients without delays are more likely to come from practices that combine data pulls with data "pushes" and other update methods — likely closer to real-time updates. The relationship between delays and specific EHR systems is particularly strong, while having multiple EHR systems is more likely among patients with no delay (Table 5).

Delayed reporting can obfuscate measurement of prevalence. Even if reporting delays are pervasive and unevenly distributed, do they affect the estimation of health outcomes? We find that using prevalence estimates that do not account for delays can significantly *distort* time-sensitive monitoring of health-related outcomes. We illustrate this phenomenon by calculating prevalence

[5] We also report relative absolute prevalence error in Section G, which is a common metric used to evaluate prevalence under class imbalance [35].

Metric	Definition	Equation
prevalence$_{(j,t)}$	Rate at which outcome Y occurs, estimated for any group j at any time t	$\frac{\sum_{i \in j} Y_i}{\sum_{i \in j} 1}$
weighted prevalence$_{(j,t)}$	Prevalence weighted by posterior probability p_{ij} when race is unobserved	$\frac{\sum_i^N p_{ij} \cdot Y_i}{\sum_i^N p_{ij}}$
prevalence error$_j$	Difference between initial and final prevalence for group j	$\text{prevalence}_{(j,t_{\text{final}})} - \text{prevalence}_{(j,t_{\text{initial}})}$
relative absolute prevalence error$_j$	Absolute difference between initial and final prevalence for group j relative to final prevalence	$\frac{\lvert \text{prevalence}_{(j,t_{\text{final}})} - \text{prevalence}_{(j,t_{\text{initial}})} \rvert}{\text{prevalence}_{(j,t_{\text{final}})}}$
average prevalence error	Absolute prevalence error averaged across all groups	$\frac{1}{G} \sum_j^G \lvert \text{ prevalence error}_j \rvert$
disparity$_{(j,j+1,t)}$	Difference in prevalence between groups j and $j+1$ at any time t	$\text{prevalence}_{(j,t)} - \text{prevalence}_{(j+1,t)}$
disparity error$_{(j,j+1)}$	Difference between initial and final disparity for two groups	$\text{disparity}_{(j,j+1,t_{\text{final}})} - \text{disparity}_{(j,j+1,t_{\text{initial}})}$
average disparity error	Absolute disparity error averaged across all pairwise group combinations	$\frac{2}{G(G-1)} \sum_j^G \sum_{j+1}^G \lvert \text{disparity error}_{(j,j+1)} \rvert$

Table 1: Error metrics for prevalence and disparity assessments from time t_{initial} (maximum number of patients with delayed reporting) to t_{final} (race fully known). Y denotes a health outcome, with $Y_i \in \{0, 1\}$ indicating the presence or absence of the outcome for an individual i at time t, in a population of N patients. j denotes an individual racial group (and $j + 1$ a different racial group), up to G total racial groups. p_{ij} denotes the posterior probability of an individual i (where $0 \leq p_{ij} \leq 1$) belonging to a specific racial group j.

	Overall Average	No Delay	Delay
N	5,310,700	1,399,487	3,911,213
Age (years)	58.02	55.97	58.76
Female (%)	56.26	55.34	56.59
Male (%)	43.69	44.58	43.37
Other (%)	4.69	7.61	3.65
AIAN (%)	0.73	0.73	0.74
Asian (%)	2.69	2.78	2.65
Black (%)	8.45	8.78	8.33
NHPI (%)	0.52	0.66	0.47
White (%)	80.25	77.49	81.24
Hisp (%)	10.91	13.52	9.98
Other (%)	4.12	4.78	3.89
Multi (%)	0.89	0.93	0.87

Table 2: On average, patients who experience delays are older and more likely to be White. We show differences in average demographic characteristics between patients who experience delays compared to patients with no delays. All differences are statistically significant ($p < 0.01$) with multiple testing corrections except AIAN (%). Also see Appendix 5.

rates for a cohort of $1,776,729$ patients from 2018 Q1. This is a subset of all $\sim$ 5.3M patients in our study, as we restrict to (1) patients whose date-of-birth is reported before 2018 (ensuring minimum patient data robustness for the time period in question) and (2) practices that report race before 2018 (eliminating practices that were not collecting race at all). We can produce estimates first at t_{initial} immediately following the conclusion of the quarter, when only 59.06% of patients have recorded race, then at regular intervals up to t_{final}, when 100% of patients have recorded race. Since our disparity assessment remains exclusively focused on 2018 Q1, health outcomes in the cohort are held fixed based on those occurring in that quarter; the only change across iterations is the proportion of patients omitted due to missing race information (which decreases over time). Table 3 summarizes changes in prevalence and disparity estimates, attributable entirely to delayed race information, between t_{final} and t_{initial} (where a number closer to zero indicates a lower error). For example, diabetes (HbA1c) testing prevalence error among Hispanic patients is 5.56 percentage points, which is higher than for other racial groups. This means that the true prevalence estimate, measured at t_{final}, is around 5 percentage points higher than the prevalence estimate at t_{initial}. Most error values are positive, consistent with our finding from Table 5 that greater delay is associated with higher prevalence of health outcomes. The average prevalence error across all groups and all outcomes is 2.15 percentage points. Figure 3 visualizes these trends at quarterly intervals from right after 2018 Q1 up to 3 years later (t_{3Yrs}), when 99.85% of patients have recorded race and the average prevalence error has been reduced to less than 0.1 percentage points. The effects of delays across multiple consecutive cohorts (2018 Q1, Q2, Q3, and Q4) are detailed in Appendix K.

We observe similar discrepancies at the state and practice level. 37 states have the same or higher amount of average prevalence error as seen at the national level, while California in particular has a lower average prevalence error (1.20 percentage points). Figure 4 illustrates state-level examples of the 2018 Q1 assessment and underscores the heterogeneity and unpredictability of delayed reporting's effects across different geographies. As shown in Figure 11 in the Appendix, most prevalence estimate errors at the practice level are small and clustered around 0, though there are a non-negligible number of outliers. Across over 1,000 individual practices, 13.39% of all prevalence estimates for each race group and health outcome are incorrect by over 10 percentage points.

Delayed reporting distorts true disparity in retrospective analyses. Do these prevalence errors affect estimates of *disparities* between demographic groups? As seen in Figures 3 and 4, even small absolute differences can lead to changes in the relative magnitude of group-level disparities. For example, an early assessment of Arkansas hypertension prevalence in 2018 Q1, conducted at t_{initial}, would lead one to conclude there is virtually no disparity between API and Hispanic patients. However, with more complete race information available after three years, one can see that the API-Hispanic hypertension gap was actually more than 5 percentage points in that quarter, with no overlap in 95% confidence intervals. Similar minimizations of disparities occur at the national level for the Hispanic-AIAN hypertension gap, the White-AIAN diabetes (HbA1c) testing gap, and the Hispanic-Black diabetes gap (see Figure 7), among others. As further detailed in Table 3 and Appendix I, the magnitude and direction of disparity error varies across pairwise comparisons, across outcomes, and across geographic scales of assessment.

Health outcome	Prevalence error					Average prevalence error	Average disparity error
	AIAN	API	Black	Hispanic	White		
Diabetes	0.88	0.86	1.12	0.26	0.75	0.81	0.57
Hypertension	1.64	0.96	2.09	0.54	1.88	1.49	1.05
Depression	1.63	0.35	0.71	0.77	1.18	0.93	0.64
Depression screen	0.97	1.25	0.87	−0.18	1.38	1.02	0.86
Electrocardiogram	1.07	5.00	4.53	3.41	4.85	3.86	2.06
HbA1c	4.14	3.60	5.26	5.56	5.34	4.78	1.34

Table 3: Errors in prevalence and disparity estimation between t_{initial} and t_{final} at the national level for each racial group and health outcome, focused on health outcomes for a cohort in 2018 Q1. Values provided are percentage points; *i.e.*, premature assessment of Hispanic HbA1c tests underestimates prevalence by 5.56 percentage points. See Table 1 for definitions of metrics and Table 6 for relative absolute prevalence error. Most prevalence errors are statistically significant ($p < 0.05$) with multiple testing correction.

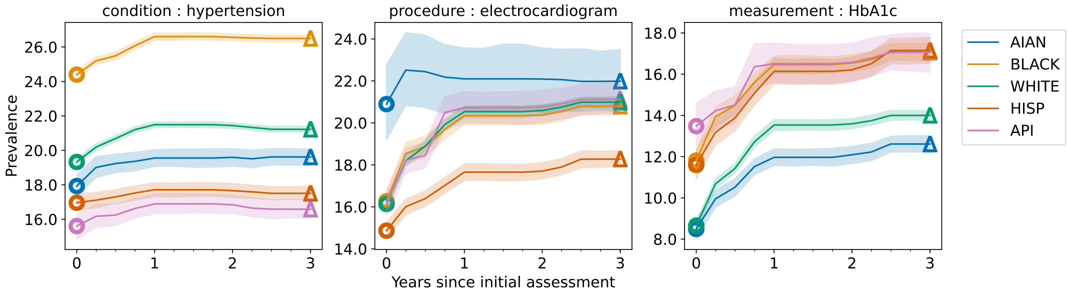

Figure 3: Simulations at the national level for one condition (hypertension), one procedure (electrocardiograms), and one measurement (HbA1c for diabetes). See additional outcomes in Figure 7. If delayed reporting had no effect on prevalence, we would expect to see horizontal lines for each race line within facets. Instead, each facet shows that rank orderings of prevalence by race changes over time, and prevalence by race often increases monotonically. Additionally, there is high uncertainty for some estimates such as electrocardiogram procedures, and those estimates experience the most fluctuation over time. All prevalence estimates are conducted on a fixed cohort of patients from 2018 Q1.

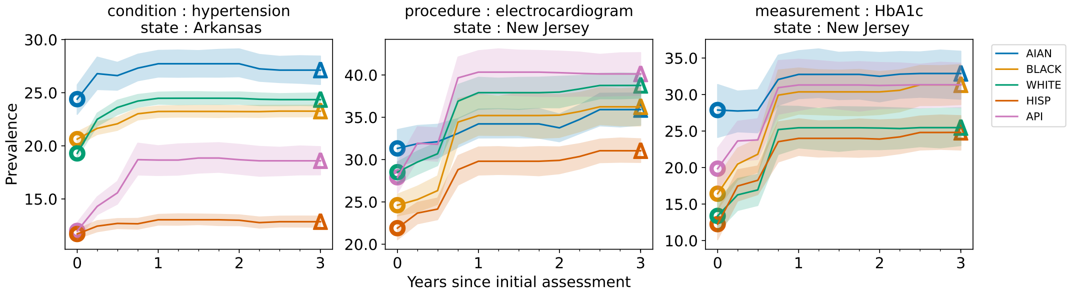

Figure 4: Simulations at the state level for hypertension diagnoses, electrocardiogram procedures, and HbA1c tests. Changes in prevalence estimates over time differ in magnitude across races, indicating variability in disparities across race groups. The states shown are each chosen from among the top three states with the highest average number of patients for each health outcome. See additional state-level outcomes in Figure 8. All prevalence estimates are conducted on a fixed cohort of patients from 2018 Q1.

Imputation is not a panacea for delayed reporting. One approach to mitigate the effect of delayed race information might be to *impute* the posterior probability of a patient's race, as is a common strategy for missing data in general. We explore whether widely used imputation methods such as Bayesian Improved First Name Surname Geocoding (BIFSG) can improve the accuracy of disparity assessments performed prior to obtaining complete race information for all patients [33, 53, 84]. Bayesian methods that predict individuals' race using a combination of first names, last names, and geography have been used across various domains to evaluate disparities [33, 34, 47, 87]. Following conventional practice [21, 34], in order to calculate a BIFSG version of prevalence, we weight each patient's contribution to each racial group's prevalence by their posterior probability p_{ij} of being in that group (see Section 4.4).

At the individual level, BIFSG achieves AUROC values > 65% for all racial groups. However, performance varies substantially across groups, with AUROCs ranging from 93.6 for Hispanic patients and

67.5 for AIAN patients,[6] suggesting group-level estimates may vary in accuracy [23]. We can see this problem in Figure 5, where prevalence estimates based on imputed probabilities from BIFSG are often over-estimated for minority groups (prevalence error is negative) though never for the majority White group. As a result of such over-estimation, BIFSG does not consistently reduce delay-based errors in disparities, which aligns with prior work on missing race or gender data [21]. We perform one-sided Mann Whitney U-tests comparing error metrics between estimates of prevalence using BIFSG and t_{3Yrs} versus $t_{initial}$ and t_{3Yrs}, with Benjamini-Hochberg correction for multiple testing. We observe that the average disparity error (detailed in Section 4.3) is only significantly improved (at 0.05 level) for diabetes diagnoses, electrocardiogram procedures, and HbA1c measurements. However, the average prevalence error does significantly decrease with the use of BIFSG for all outcomes, though the size of the difference is numerically small for some outcomes (see Figure 12 in the Appendix). Note that results are sensitive to rounding of p_{ij}. These results indicate that imputation methods like BIFSG can mitigate delayed reporting to some degree with regard to prevalence errors, but do not produce accurate prevalence point estimates nor disparity error estimates. Thus, BIFSG cannot always replace accurate, self-reported race information in prevalence and disparity estimation.

6 Discussion: Implications for Practitioners

Our findings inform how to improve the use of data-driven decision-making tools in light of demographic reporting delays.

More holistic efforts should be made to understand and address the mechanisms driving delays. While there is an extensive literature on accounting for and imputing missing data in healthcare, the impact of changing missingness over time is less studied. Thus, modeling missingness mechanisms is an interesting direction of future work. In our case, this might involve efforts to uncover why delayed reporting is correlated with health outcomes, and the precise mechanisms through which delays occur during the patient intake and reporting process. This recommendation touches on a key limitation of our study, which is a retrospective analysis of a de-identified, pre-existing dataset. Given that we do not have the ability to contact the individual decision-makers (*e.g.*, nurses, clinicians, intake coordinators, etc.), we are unable to explore the many upstream factors that may have caused delayed race reporting. More qualitative analyses are essential to uncover the drivers of delayed reporting. For example, surveys could be conducted across practice sites to understand common data collection protocol and infrastructural reasons for delays.

One implication of our findings is that different EHR systems vary in their patient delayed reporting rates, suggesting that user design choices, as well as backend software architectures, may contribute to delays — a meaningful overlap between human computer interaction (HCI) and fairness domains. Since rates of delay also vary across racial groups, these efforts should further consider the role of individual behavior — such as hesitance to report race — and practice-side variation in recordkeeping. We build on calls in the social sciences for incorporating qualitative research methods in data science work [42], and complement existing efforts to document and understand practitioner experiences with data collection [7], with appropriate data protection mechanisms (and communication to patients thereof) [58]. This recommendation also aligns with the growing recognition in algorithmic fairness that decision-making tools should be studied in their institutional contexts [83, 86].

Understanding delays retrospectively is complex and challenging to test. Although we cannot study the precise mechanisms of delays, we hypothesize several mechanisms through which delays might occur, and why there are higher rates of delay among White patients and patients with more health conditions.

Patient hesitance: While we cannot directly measure hesitance, we study patients whose reported race changes from "declined" or "unknown" to "known." We find only small differences in the proportion of White patients in this group relative to other patients in the data, which suggests that hesitance does not play a significant role in the delays that we observe here.

Systemic Complexity and Delayed Presentation: Patients with complex health conditions might have more complicated medical records and/or more administrative tasks [60], thus potentially leading to administrative delays in fully completing demographic information [73].

Intake / registration visits: Some patient visits might be intake visits, where race data may be missing when these are not properly administered. Prior research suggests that Black and Hispanic patients are more likely to utilize the emergency department [9, 51] and may face greater barriers to accessing primary care regularly [19]. Racial minorities also have lower health insurance rates [48], an additional barrier to scheduling routine, preventative care visits. To evaluate the role of intake visits, we test whether removing the first timestamp associated with a patient's DOB would eliminate reporting delays. If intake visits explained reporting delays, we would expect race information to be recorded by the second timestamp. While we observe a significant decrease in the percentage of patients with delays (around 20 percentage points), it does not appear that intake visits explain all of the reporting delays in our data.

Electronic health record (EHR) system-dependent lags: Patients whose race is collected from practices with specific EHR systems might experience greater lags (*e.g.*, due to specific data entry workflows). Prior research has shown tradeoffs between verbally collecting information from patients compared to paper forms or tablets [82]. Table 5 provides some evidence of EHR-system differences.

Several features of our study may also limit the generalizability of some findings. We focus only on a primary healthcare setting, and do not include data from other external databases. Conclusions from research on hospitals and emergency departments may be less applicable. As mentioned earlier, the dataset and cohort used in our analyses were retrospective. In particular, the dataset was not specifically constructed to be a representative sample of the U.S. population, despite the AFC dataset having relatively representative coverage. Lastly, our study is limited to U.S. health data. Findings may differ in other geographic settings, or with other notions of social identity such as caste [78].

Imputation alone is not enough; practitioners should invest in improving data collection efforts on the ground. We

[6] We note that BIFSG AUROC for patients in 'Other' race groups was low (34.22). For fair comparison with BIFSG, we exclude this group from all simulations.

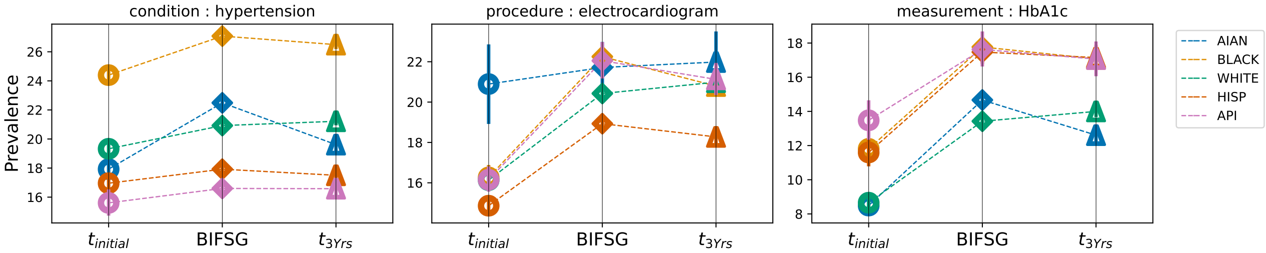

Figure 5: BIFSG over-estimates prevalence for several minority race groups (*e.g.*, Black and AIAN patients) across several outcomes, though average prevalence error is improved. For example, prevalence estimates with BIFSG for Black patients (yellow diamonds) are higher than estimates at $t_{initial}$ (yellow circles) and t_{3Yrs} (yellow triangles), but are closer to estimates at t_{3Yrs} (yellow triangles). In each subplot, the y-axis denotes the estimated prevalence. Values for $t_{initial}$ and t_{3Yrs} match the same national values as shown in Figure 3. See additional outcomes in Figure 14.

find that a widely used imputation method may perform satisfactorily on individual-level accuracy, but does not fundamentally improve group-level disparity assessments affected by reporting delays. In situations where timely feedback is necessary, our findings suggest practitioners should advocate for the suite of policy and programmatic changes that may reduce delays in the first place. It is important to consider principles such as data minimization [14, 58], and ensure robust privacy protections [71] when designing strategies to incentivize timely data reporting. Augmenting existing data collection with more reliable demographic data sources may be another promising direction — *e.g.*, integrating EHR and insurance data, which may have higher reporting standards.

Future work should study the impact of delays on other types of disparity metrics and ML-based metrics. Importantly, outcomes of interest may not be binary (*e.g.*, number of days to readmission after discharge from a hospital). Similarly, the impact of delayed race reporting should also be assessed in the context of fairness metrics corresponding to ML-based predictive models (*e.g.*, models predicting clinical interventions such as vassopressor administration [40]). In this vein, our work connects to the literature on algorithmic audits in ML [61, 77] — in particular, how missing and unreliable demographic data may impede auditing efforts [10]. Our work suggests one reason why static audits – completed once in time – may fail to detect disparities. Furthermore, when conducting disparity assessments that involve aggregating data across sites or practices, it may be important to design data sampling and non-respondent follow-up strategies that account for delayed reporting.

Lastly, it's worth noting that we exclude patients (approximately 11%) for whom race information is *never available*. Even though we are ultimately able to recover race data for all of the patients in our cohort — those with reporting *delays* — traditional sources of missingness may also bias our disparity estimates.

The existence of delays in race reporting underscores the importance of continuously and dynamically assessing fairness. Our results show that rates of missingness can be different between groups at different points of time. Hence, continuous assessment would be required to assess the robustness of conclusions made about disparities. It is important to consider sociotechnical systems as *dynamic*, and how data missingness rates may change over time, driven by repeated user interactions with the same system (see Section K in the Appendix, where we analyze delayed reporting for consecutive patient cohorts and discuss implications for real-time monitoring scenarios). Prior work in algorithmic fairness has similarly raised the issue that fairness research should study the long-term impacts of deployed systems [28]. Our work aligns with such concerns, though we focus on underlying changes in the data. In particular, we urge fairness researchers to avoid treating data inputs as fixed, and to re-evaluate historical disparity assessments as more data that was initially missing becomes available over time. Our work also suggests another vulnerability to current static audit approaches: in the presence of reporting delays, providers might advertently or inadvertently leverage reporting delays to achieve more favorable audit outcomes.

7 Conclusion

In this work, we demonstrate the impact of delayed demographic information reporting when auditing the fairness of decision-making systems. We focus on applications to healthcare where regular and timely monitoring of health disparities is critical. However, our work extends to any setting in which time-sensitive evaluations must be conducted prematurely. In a nationwide health dataset, we find that delayed reporting is a widespread problem, affecting nearly 3 out of every 4 patients. Furthermore, delays do not impact all patients evenly. Rates of delayed reporting vary by race and there are demographic, health, and practice-level differences between patients with and without delays. Furthermore, when we retrospectively estimate the impact of delays on disparity assessments, we find that delays can lead to inaccurate depictions of disparities. While these distortions are relatively small at the national level, there is greater heterogeneity for estimates at the state and practice levels — an important consideration as recent health equity initiatives have occurred at the state level, and mitigation efforts necessarily start at the practice level.

Broadly, our work highlights a crucial gap in the current auditing space: the need for frequent monitoring of systems reliant on seemingly "static" variables like race. In fast-paced deployment environments, delays in specific data inputs may arise, leading to unexpected performance. Prior research has pointed out that access to individual-level demographic data is often unrealistic in real-world settings [7, 8, 10, 50, 62]. Our work complicates this finding: demographic data can also be *delayed*, thus highlighting an important direction for future work.

8 Ethical Statement

We now address ethical considerations that arose in the course of this work.

First, our work deals with sensitive patient health data. All analyses were conducted on secure, remote servers approved for High Risk and Protected Health Information (PHI) data, and the research was approved by the Institutional Review Board, including Waiver of Informed Consent, Waiver of Assent, and Waiver of HIPAA Authorization. To protect confidentiality, we only produced aggregated results for cell sizes > 10. The American Family Cohort dataset is used solely for research purposes, allowing researchers to investigate core questions of health equity and to generate knowledge that may inform the improvement of healthcare services across a nationwide network of primary care practices.

Second, another ethical consideration lies in the measurement of race. Through all of our findings, our central focus is on the sobering reality that health disparities exist between different groups within communities across the U.S. The true nature of these disparities are, of course, always more complex and intersectional than the socially constructed racial categories we choose to use at any given moment (*i.e.*, individuals' self-reported racial categories may not align with federal categories or may change over time), and the improper reification of racial categories (including through statistical imputation methods such as BIFSG) may itself run the risk of feeding back into the entrenchment or exacerbation of those disparities. At the same time, in the absence of any records of race identification, we may not be able to detect and act upon disparities at all. Central to algorithmic fairness has been the notion of "fairness through awareness" [31]. At core, our study identifies an underappreciated mechanism, delayed demographic reporting, by which that awareness can be obfuscated. Critically, the utilization of racial categories, obtained through self-reporting or through imputation methods, to *assess disparities* does not in any way imply that they should also be used in individual-level medical decision making [49], where the ethical considerations may be significantly more acute.

Acknowledgments

We thank Nathaniel Hendrix, Shiying Hao, Esther Velásquez, Ayin Vala, and Isabella Chu for their support with using American Family Cohort data, and Julian Morimoto, Claire Morton, and Jenn Wang for their helpful feedback. We are grateful to the Stanford Institute for Human-Centered Artificial Intelligence and the Robert Wood Johnson Foundation Evidence 4 Action Program for supporting this research.

References

[1] 2021. Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and

[2] Hammaad Adam, Fan Yin, Huibin Hu, Neil Tenenholtz, Lorin Crawford, Lester Mackey, and Allison Koenecke. 2023. Should I stop or should I go: early stopping with heterogeneous populations. *Advances in Neural Information Processing Systems* 36 (2023), 15799–15832.

[3] Nil-Jana Akpinar, Zachary Lipton, and Alexandra Chouldechova. 2024. The Impact of Differential Feature Under-reporting on Algorithmic Fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* 1355–1382.

[4] Nil-Jana Akpinar, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, and Hoda Heidari. 2022. A sandbox tool to bias (stress)-test fairness algorithms. *arXiv preprint arXiv:2204.10233* (2022).

[5] Kelli D Allen, Eugene Z Oddone, Cynthia J Coffman, Francis J Keefe, Jennifer H Lindquist, and Hayden B Bosworth. 2010. Racial differences in osteoarthritis pain and function: potential explanatory factors. *Osteoarthritis and cartilage* 18, 2 (2010), 160–167.

[6] Karen O Anderson, Carmen R Green, and Richard Payne. 2009. Racial and ethnic disparities in pain: causes and consequences of unequal care. *The Journal of Pain* 10, 12 (2009), 1187–1204.

[7] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 249–260.

[8] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 1709–1721.

[9] MJ Arnett, Roland J Thorpe, DJ Gaskin, Janice V Bowie, and Thomas A LaVeist. 2016. Race, medical mistrust, and segregation in primary care as usual source of care: findings from the exploring health disparities in integrated communities study. *Journal of Urban Health* 93 (2016), 456–467.

[10] Carolyn Ashurst and Adrian Weller. 2023. Fairness without demographic data: A survey of approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–12.

[11] American Medical Association. [n. d.]. How to select a practice management system. https://www.ama-assn.org/practice-management/claims-processing/how-select-practice-management-system

[12] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 206–214.

[13] Brett Beaulieu-Jones, Samuel G Finlayson, Corey Chivers, Irene Chen, Matthew McDermott, Jaz Kandola, Adrian V Dalca, Andrew Beam, Madalina Fiterau, and Tristan Naumann. 2019. Trends and focus of machine learning applications for health research. *JAMA network open* 2, 10 (2019), e1914051–e1914051.

[14] Asia J Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. 2020. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval.* 399–408.

[15] Arlene S Bierman, Nicole Lurie, Karen Scott Collins, and John M Eisenberg. 2002. Addressing racial and ethnic barriers to effective health care: the need for better data. *Health Affairs* 21, 3 (2002), 91–102.

[16] Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho, and Hoda Heidari. 2023. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–11.

[17] D Keith Branham, Kenneth Finegold, Lucy Chen, Melony Sorbero, Roald Euller, Marc N Elliott, and Benjamin D Sommers. 2022. Trends in missing race and ethnicity information after imputation in HealthCare. gov marketplace enrollment data, 2015-2021. *JAMA Network Open* 5, 6 (2022), e2216715–e2216715.

[18] Karla L Caballero Barajas and Ram Akella. 2015. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 69–78.

[19] César Caraballo, Chima D Ndumele, Brita Roy, Yuan Lu, Carley Riley, Jeph Herrin, and Harlan M Krumholz. 2022. Trends in racial and ethnic disparities in barriers to timely medical care among adults in the US, 1999 to 2018. In *JAMA Health Forum*, Vol. 3. American Medical Association, e223856–e223856.

[20] Jerome H Carter. 2008. *Electronic health records: a guide for clinicians and administrators.* ACP Press.

[21] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* Association for Computing Machinery, New York, NY, USA, 339–348. doi:10.1145/3287560.3287594

[22] Lingwei Cheng, Isabel O Gallegos, Derek Ouyang, Jacob Goldin, and Dan Ho. 2023. How redundant are redundant encodings? blindness in the wild and racial disparity when race is unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 667–686.

[23] Sergey Chernenko and David S Scharfstein. 2023. The Limits of Algorithmic Measures of Race in Studies of Outcome Disparities. *Available at SSRN 4426161* (2023).

[24] Isabel Chien, Nina Deliu, Richard Turner, Adrian Weller, Sofia Villar, and Niki Kilbertus. 2022. Multi-disciplinary fairness considerations in machine learning for clinical trials. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 906–924.

[25] CMS. 2018. MEASURES MANAGEMENT SYSTEM. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-instruments/MMS/Downloads/A-Brief-Overview-of-Qualified-Clinical-Data-Registries.pdf

[26] CMS. 2025. CMS Framework for Health Equity. https://www.cms.gov/priorities/health-equity/minority-health/equity-programs/framework.

[27] National Research Council et al. 2004. Eliminating health disparities: Measurement and data needs. (2004).

[28] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[29] Jacob W Dembosky, Amelia M Haviland, Ann Haas, Katrin Hambarsoomian, Robert Weech-Maldonado, Shondelle M Wilson-Frederick, Sarah Gaillot, and Marc N Elliott. 2019. Indirect estimation of race/ethnicity for survey respondents who do not report race/ethnicity. *Medical care* 57, 5 (2019), e28–e33.

[30] Stephen F Derose, Richard Contreras, Karen J Coleman, Corinna Koebnick, and Steven J Jacobsen. 2013. Race and ethnicity data quality and imputation using US Census data in an integrated health system: the Kaiser Permanente Southern California experience. *Medical Care Research and Review* 70, 3 (2013), 330–345.

[31] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[32] Laura Dwyer-Lindgren, Parkes Kendrick, Yekaterina O Kelly, Dillon O Sylte, Chris Schmidt, Brigette F Blacker, Farah Daoud, Amal A Abdi, Mathew Baumann, Farah Mouhanna, et al. 2022. Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities. *The Lancet* 400, 10345 (2022), 25–38.

[33] Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* 9 (2009), 69–83.

[34] Hadi Elzayn, Evelyn Smith, Thomas Hertz, Cameron Guage, Arun Ramesh, Robin Fisher, Daniel E Ho, and Jacob Goldin. 2024. Measuring and mitigating racial disparities in tax audits. *The Quarterly Journal of Economics* (2024), qjae027.

[35] Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. 2023. *Learning to quantify*. Springer Nature.

[36] Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems* 36, 7 (2021), 3217–3258.

[37] Agency for Healthcare Research and Quality. 2024. 2023 National Healthcare Quality and Disparities Report. https://www.ahrq.gov/research/findings/nhqrdr/nhqdr23/index.html

[38] Allen Fremont, Joel S Weissman, Emily Hoch, and Marc N Elliott. 2016. When race/ethnicity data are lacking: using advanced indirect estimation methods to measure disparities. *RAND Health Quarterly* 6, 1 (2016).

[39] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

[40] Marzyeh Ghassemi, Marco Pimentel, Tristan Naumann, Thomas Brennan, David Clifton, Peter Szolovits, and Mengling Feng. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[41] Carmen R Green, Karen O Anderson, Tamara A Baker, Lisa C Campbell, Sheila Decker, Roger B Fillingim, Donna A Kaloukalani, Kathyrn E Lasch, Cynthia Myers, Raymond C Tait, et al. 2003. The unequal burden of pain: confronting racial and ethnic disparities in pain. *Pain Medicine* 4, 3 (2003), 277–294.

[42] Nikolitsa Grigoropoulou and Mario L Small. 2022. The data revolution in social science needs qualitative research. *Nature Human Behaviour* 6, 7 (2022), 904–906.

[43] Hyeouk Chris Hahm, Benjamin Le Cook, Andrea Ault-Brutus, and Margarita Alegría. 2015. Intersection of race-ethnicity and gender in depression care: screening, access, and minimally adequate treatment. *Psychiatric Services* 66, 3 (2015), 258–264.

[44] Fern R Hauck, Kawai O Tanabe, and Rachel Y Moon. 2011. Racial and ethnic disparities in infant mortality. In *Seminars in perinatology*, Vol. 35. Elsevier, 209–220.

[45] J Sonya Haw, Megha Shah, Sara Turbow, Michelle Egeolu, and Guillermo Umpierrez. 2021. Diabetes complications in racial and ethnic minority populations in the USA. *Current Diabetes Reports* 21 (2021), 1–8.

[46] Margaret M Heckler. 1985. Report of the Secretary's Task Force Report on Black and Minority Health: The Heckler Report.

[47] Peter Hepburn, Renee Louis, and Matthew Desmond. 2020. Racial and gender disparities among evicted Americans. *Sociological Science* 7 (2020), 649–662.

[48] Latoya Hill, Nambi Ndugga, Samantha Artiga, and Anthony Damico. 2025. Health Coverage by Race and Ethnicity, 2010-2023. https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-coverage-by-race-and-ethnicity/

[49] Daniel E Ho and Alice Xiang. 2020. Affirmative algorithms: The legal grounds for fairness as awareness. *U. Chi. L. Rev. Online* (2020), 134.

[50] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[51] Rick Hong, Brigitte M Baumann, and Edwin D Boudreaux. 2007. The emergency department for routine healthcare: race/ethnicity, socioeconomic status, and perceptual factors. *The Journal of Emergency Medicine* 32, 2 (2007), 149–158.

[52] Elizabeth A Howell. 2018. Reducing disparities in severe maternal morbidity and mortality. *Clinical Obstetrics and Gynecology* 61, 2 (2018), 387–399.

[53] Kosuke Imai and Kabir Khanna. 2016. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis* 24, 2 (2016), 263–272.

[54] Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. 2003. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press (US), Washington (DC). http://www.ncbi.nlm.nih.gov/books/NBK220358/

[55] Cara V James, Jennifer M Haley, Eva H Allen, and Taylor Nelson. 2023. Using Race and Ethnicity Data to Advance Health Equity. https://www.urban.org/research/publication/using-race-and-ethnicity-data-advance-health-equity

[56] Vincent Jeanselme, Maria De-Arteaga, Zhe Zhang, Jessica Barrett, and Brian Tom. 2022. Imputation strategies under clinical presence: Impact on algorithmic fairness. In *Machine Learning for Health*. PMLR, 12–34.

[57] Falaah Arif Khan, Denys Herasymuk, Nazar Protsiv, and Julia Stoyanovich. 2024. Still More Shades of Null: A Benchmark for Responsible Missing Value Imputation. *arXiv preprint arXiv:2409.07510* (2024).

[58] Jennifer King, Daniel Ho, Arushi Gupta, Victor Wu, and Helen Webley-Brown. 2023. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in US Government. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 492–505.

[59] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.

[60] Michael Anne Kyle and Austin B Frakt. 2021. Patient administrative burden in the US health care system. *Health Services Research* 56, 5 (2021), 755–765.

[61] Khoa Lam, Benjamin Lange, Borhane Blili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. 2024. A framework for assurance audits of algorithmic systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1078–1092.

[62] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[63] Dingwen Li, Patrick Lyons, Jeff Klaus, Brian Gage, Marin Kollef, and Chenyang Lu. 2021. Integrating static and time-series data in deep recurrent models for oncology early warning systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 913–936.

[64] Roderick JA Little and Donald B Rubin. 1989. The analysis of social science data with missing values. *Sociological Methods & Research* 18, 2-3 (1989), 292–326.

[65] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

[66] Marian F MacDorman, Marie Thoma, Eugene Declcerq, and Elizabeth A Howell. 2021. Racial and ethnic disparities in maternal mortality in the United States using enhanced vital records, 2016–2017. *American Journal of Public Health* 111, 9 (2021), 1673–1681.

[67] David Machledt. 2021. Addressing health equity in Medicaid managed care. *National Health Law Program* (2021).

[68] Susan E Manning, Antonia M Blinn, Sabrina C Selk, Christine F Silva, Katie Stetler, Sarah L Stone, Mahsa M Yazdy, and Monica Bharel. 2022. The Massachusetts racial equity data road map: data as a tool toward ending structural racism. *Journal of Public Health Management and Practice* 28, Supplement 1 (2022), S58–S65.

[69] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163.

[70] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 429–438.

[71] The National Artificial Intelligence Advisory Committee (NAIAC). 2024. RECOMMENDATION: Data Challenges and Privacy Protections for Safeguarding Civil Rights in Government. https://ai.gov/wp-content/uploads/2024/06/RECOMMENDATION_Data-Challenges-and-Privacy-Protections-for-Safeguarding-Civil-Rights-in-Government.pdf

[72] Engineering National Academies of Sciences and Medicine. 2017. *Communities in Action: Pathways to Health Equity*. The National Academies Press, Washington, DC. doi:10.17226/24624

[73] National Association of Community Health Centers. 2023. Closing the Primary Care Gap.

[74] California Department of Health Care Access and Information. 2022. Hospital Equity Measures Reporting Program. https://hcai.ca.gov/data/healthcare-quality/hospital-equity-measures-reporting-program/

[75] U.S. Office of Management and Budget. 2024. 1997 Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. https://spd15revision.gov/content/spd15revision/en/history/1997-standards.html

[76] Fernanda CG Polubriaginof, Patrick Ryan, Hojjat Salmasian, Andrea Wells Shapiro, Adler Perotte, Monika M Safford, George Hripcsak, Shaun Smith, Nicholas P Tatonetti, and David K Vawdrey. 2019. Challenges with quality of race and ethnicity data in observational databases. *Journal of the American Medical Informatics Association* 26, 8-9 (2019), 730–736.

[77] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[78] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 315–328.

[79] Keith R Spangler, Jonathan I Levy, M Patricia Fabian, Beth M Haley, Fei Carnes, Prasad Patil, Koen Tieskens, R Monina Klevens, Elizabeth A Erdman, T Scott Troppy, et al. 2023. Missing race and ethnicity data among COVID-19 cases in Massachusetts. *Journal of Racial and Ethnic Health Disparities* 10, 4 (2023), 2071–2080.

[80] Center for Population Health Sciences Stanford Medicine. [n. d.]. American Family Cohort (AFC).

[81] Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.

[82] Health Research & Educational Trust. 2013. Reducing Health Care Disparities: Collection and Use of Race, Ethnicity and Language Data. https://www.aha.org/system/files/hpoe/Reports-HPOE/equity-care-report-august2013.PDF

[83] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[84] Ioan Voicu. 2018. Using first name information to improve race and ethnicity classification. *Statistics and Public Policy* 5, 1 (2018), 1–13.

[85] Sidney D Watson. 2024. Community Engagement, Public Reporting, and Financial Incentives: Lessons from Michigan on Tackling Racial and Ethnic Disparities in Medicaid Managed Care. *Houston Journal of Health Law & Policy* 23, 1 (2024), 111–144.

[86] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4477–4488.

[87] Yan Zhang. 2018. Assessing fair lending risks using race/ethnicity proxies. *Management Science* 64, 1 (2018), 178–197.

[88] Yiliang Zhang and Qi Long. 2021. Assessing fairness in the presence of missing data. *Advances in Neural Information Processing Systems* 34 (2021), 16007–16019.

[89] James Zou, Judy Wawira Gichoya, Daniel E Ho, and Ziad Obermeyer. 2023. Implications of predicting race variables from medical images. *Science* 381, 6654 (2023), 149–150.

A Detecting when Race and Ethnicity is Unknown or Declined

The dataset contains multiple datapoints (at different timepoints) per patient / practice. Each datapoint is associated with a modification date of race. Further, each datapoint contains two free-text fields – 'patientracetext' and 'patientethnicitytext'. Race codes are also available in fields of 'patientracecode' and 'patientraceethnicity'. We define patient race to be unknown or declined at a given timepoint if both the following conditions are met:

(1) Free text fields (after lower-casing) are in the following list: 'race not reported - don't know', 'nh', 'unspecified', 'do not use', 'not hispanic/latino ethnicity', 'other/declined', 'non-hispanic', 'not of hispanic, latino/a or spanish origin', 'refuse', 'not hispanic', 'declined', 'unknown', 'patient declined information', 'unreported/refused to report', 'patient declined', 'not set', 'refuse to report/ unreported', 'declined to answer', 'not reported', 'non hispanic', 'withheld', 'unknown/unwilling', 'u', '<none>', 'NA', 'unknown/unreported', 'decline to answer', 'refused to report', 'unknown / not reported', 'not hispanic or latino', 'non hispanic-non latino', 'non - hispanic/latino', 'prefers not to answer', '*unspecified', 'refused', 'refused to report/unreported', 'unknown/not reported', 'not hispanic / latino', 'dec', 'refused, unknown', 'undefined', 'chose not to disclose', 'unk', 'race not reported - refusal', 'non hispanic or latino', 'n', 'nsp', 'x', 'unk', 'declines to state', 'unavailable / unknown', 'refus', 'dec', 'not hispanic, latino/a, or spanish origin', 'non-hispanic / non latino', 'state prohibited', 'decline', 'declined to specify', 'not provided', 'patient refused', 'un', 'unreported / refused to report', 'race not reported - not ascertained', 'unknown to patient', 'declines to specify', 'decli', 'dc', 'ds', 'ua', 'uo', 'n', 'd', 'u', 'r', 'unkno', 'nr', 'unreported / unknown (uds)', 'unreported / unknown', 'unavailable', '2186-5', '9'.

(2) Categorical codes are either invalid codes or strings indicating no information. Specifically, they are among the following list: 'UNK', 'NA', 'UN', 'U', '2186-5', nan, 'UNK', 'UN', '2186 - 5', 'N', '312507', 'NH', 'NR', 'ASKU'.

Note that if some information is provided in either field – race or ethnicity – we do not consider race to be unreported. The only exception is when patients only report that they are non-Hispanic, with no other race information provided. In such cases, race is considered missing or delayed.

B Extraction of Health Outcomes

We extract six clinical outcomes curated based a literature review: *i.e.*, there is evidence of disparities between racial group for each of these outcomes. For example, prior work has documented higher depression screening rates among Black and Asian patients compared to White patients [43]. Health outcomes in each case are extracted by relying primarily on Systematized Nomenclature of Medicine (SNOMED) codes, which are used for clinical documentation and billing purposes. Codes in each case are retrieved using a database search tool. We extract health outcomes based on the presence of specific diagnosis, documentation, and billing codes. We rely primarily on SNOMED-CT codes because they are also used for clinical documentation, apart from just billing. To identify relevant codes per outcome, we use the Athena search tool.[7] Using a relevant search term per outcome, we retrieve a list of codes that are returned as matches per search. Then, we filter these codes manually by reading the text description corresponding to the code. We assume that a code, if entered in the system, is accurate. An overview of the codes for each outcome is provided in Table 4. To fully account for outcomes, we also include codes where the presence of an outcome is indirectly indicated. For example, the SNOMED-CT code corresponding to the condition of "Senile dementia with depression" is also included in the list of codes for identifying the outcome of depression diagnosis.

Clinical Outcome	Vocabulary	Number of codes (example)
Depression diagnosis	SNOMED, OMOP Extension	104 (e.g., 35489007)
Diabetes diagnosis	SNOMED, OMOP Extension	92 (e.g., 771000119108)
Hypertension diagnosis	SNOMED, OMOP Extension	10 (e.g., 78975002)
Diabetes HBA1c measurement	SNOMED, OMOP Extension	1 (43396009)
Electrocardiogram procedure	SNOMED, CPT4, HCPCS	38 (e.g., 93005)
Depression screening procedure	SNOMED, CPT4, HCPCS	4 (e.g., 96127)

Table 4: Overview of codes for each outcome variable.

C Cohort Size versus Delay

In Figure 6, we visualize cohort size versus average delay (in days), as they vary across different quarters for which to conduct the disparity assessment. We choose the 2018 Q1 cohort because it has a reasonable sample size, as well as a high average delay.

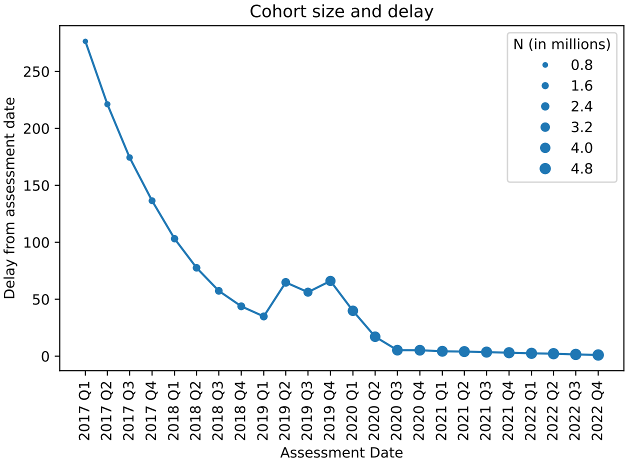

Figure 6: Cohort size vs average delay of race reporting for different quarters as the focus of disparity assessment.

D Comparing Patients with and without Delays

Table 5 shows the average difference in demographic characteristics and all other variables considered in the main text.

E Retrospective Analysis for Additional Health Outcomes

Figure 7 and Figure 8 show the results from the retrospective analysis at both the national and state level for three additional health outcomes: diabetes, depression, and depression screens.

[7] https://athena.ohdsi.org/search-terms/start

		Overall Average	**No Delay**	**Delay**	**Difference**
	N	5,310,700	1,399,487	3,911,213	
	Patient-level Characteristics				
Demographics	Age (years)	58.02	55.97	58.76	2.79^{***}
	Female (%)	56.26	55.34	56.59	1.25^{***}
	Male (%)	43.69	44.58	43.37	-1.21^{***}
	Other (%)	4.69	7.61	3.65	-3.96^{***}
	AIAN (%)	0.73	0.73	0.74	0.01
	Asian (%)	2.69	2.78	2.65	-0.12^{***}
	Black (%)	8.45	8.78	8.33	-0.45^{***}
	NHPI (%)	0.52	0.66	0.47	-0.18^{***}
	White (%)	80.25	77.49	81.24	3.75^{***}
	Hisp (%)	10.91	13.52	9.98	-3.54^{***}
	Other (%)	4.12	4.78	3.89	-0.89^{***}
	Multi (%)	0.89	0.93	0.87	-0.07^{***}
Marital Status	Single (%)	28.28	30.26	27.64	-2.62^{***}
	Married (%)	58.72	57.26	59.19	1.93^{***}
	Divorced (%)	7.42	7.33	7.45	0.12^{***}
	Widowed (%)	5.51	5.02	5.67	0.65^{***}
	Partner (%)	0.07	0.12	0.05	-0.08^{***}
	Other (%)	4.69	7.61	3.65	-3.96^{***}
Visits	Avg. yearly visits (2007-2023)	0.82	0.50	0.93	0.43^{***}
	Earliest year	2016.23	2017.08	2015.92	-1.15^{***}
Health	Diabetes (%)	10.84	8.54	11.66	3.12^{***}
	Depression (%)	12.33	9.90	13.20	3.30^{***}
	Hypertension (%)	25.17	20.24	26.93	6.70^{***}
	Depression screen (%)	14.44	10.26	15.94	5.67^{***}
	Electrocardiogram (%)	26.84	18.63	29.78	11.15^{***}
	Hba1c (%)	25.46	17.74	28.22	10.49^{***}
	Practice-level Characteristics (Mapped to Patients)				
Practice Info.	Available (%)	85.93	86.25	85.82	-0.43^{***}
	Unavailable (%)	14.07	13.75	14.18	0.43^{***}
Data Source	EHR & PM (%)	6.38	9.86	5.13	-4.73^{***}
	EHR only (%)	93.02	89.59	94.25	4.66^{***}
Data Update	Push, pull, and other (%)	17.43	21.87	15.84	-6.03^{***}
	Pull only (%)	81.07	76.39	82.76	6.37^{***}
	Other (%)	1.45	1.70	1.36	-0.34^{***}
EHR System	Multiple (%)	17.71	22.16	16.11	-6.05^{***}
	eMDs - Solution Series (%)	51.96	45.19	54.40	9.21^{***}
	Amazing Charts (%)	8.60	9.22	8.38	-0.84^{***}
	eMDs - Practice Partner (%)	3.18	4.43	2.73	-1.70^{***}
	Veradigm EHR (%)	3.00	3.67	2.76	-0.91^{***}
	eClinicalWorks (%)	1.82	1.12	2.07	0.95^{***}
	GE Centricity (%)	1.75	1.20	1.95	0.75^{***}
	Aprima (%)	1.63	1.55	1.66	0.11^{***}
	Athenahealth (%)	1.33	2.06	1.07	-0.99^{***}
	eMDs - Lytec MD (%)	1.21	2.02	0.91	-1.11^{***}
	Other (%)	7.22	7.00	7.30	0.30^{***}

Table 5: Differences in average demographic characteristics, visits, health outcomes, and practice-level characteristics between patients who experience delays compared to patients with no delays. Most of the differences are statistically significant with $* * * = p < 0.01$ even with Benjamini-Hochberg adjustment for multiple comparisons. On average, patients who experience delays are older, more likely to be White, and have more visits. They also have higher prevalence rates of adverse health conditions and procedures. AIAN = American Indian or Alaska Native; NHPI = Native Hawaiian or Pacific Islander; EHR = electronic health record; PM = practice management system.

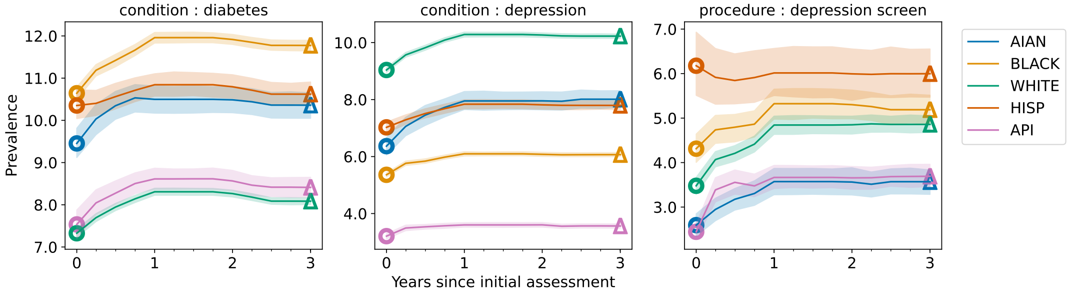

Figure 7: Simulations at the national level for two conditions (diabetes and depression) and one procedure (depression screens). All disparity estimates are conducted on a fixed cohort of patients from 2018 Q1.

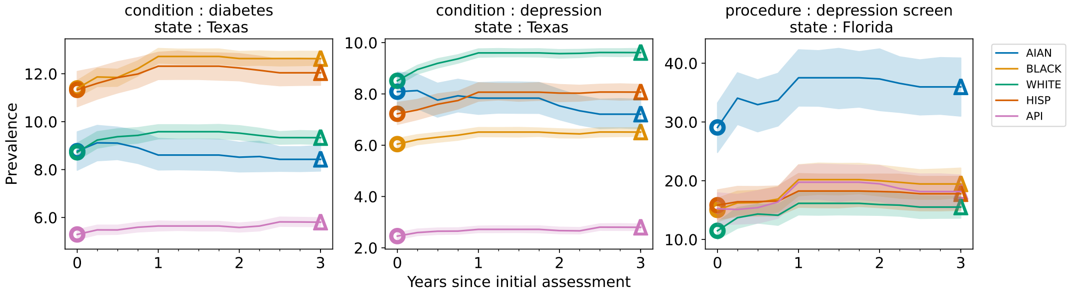

Figure 8: Simulations at the state level for two conditions (diabetes and depression) and one procedure (depression screens). The states shown are among the top three states with the highest average number of patients with each health outcome. All disparity estimates are conducted on a fixed cohort of patients from 2018 Q1.

F Delayed Reporting for 2017 Q1 and 2018 Q1 Cohorts

Figures 9 and 10 implement national-level disparity assessments for different cohorts—2017 Q1 and 2019 Q1. They show temporal differences in the effect of delays. In 2019 Q1, delays play a minor role—affecting prevalence and disparity estimates slightly (*e.g.*, the Black-API disparity for HbA1c measurements decreases in $t_{3\text{Yrs}}$) compared to t_{initial}). Delays have a much larger effect on disparity estimates for the 2017 Q1 cohort in line with the findings from Figure 6 that show greater average delay in 2017 compared to 2019 and later.

G Relative Absolute Prevalence Error

We show relative absolute prevalence error by race in Table 6. Relative absolute error accounts for class imbalance, and shows the magnitude of the prevalence error relative to the true overall prevalence. For example, prevalence errors for hypertension and depression are broadly similar among AIAN, Hispanic, and White subgroups in Table 3. But relative absolute errors (shown here) are much larger for depression compared to hypertension since the true overall prevalence of depression overall is smaller.

H Distribution of Errors in Prevalence

Figure 11 shows the distribution of prevalence errors at the practice level across different racial groups and for all health outcomes. We present prevalence errors averaged across 50 bootstrapped samples. Most of the errors are centered around 0, though there are numerous outliers. Overall, prevalence errors skew positive, reflecting a trend similar to the national level (Table 3).

Health outcome	Relative absolute prevalence error				
	AIAN	API	Black	Hispanic	White
Diabetes	8.79	10.79	9.54	3.92	9.34
Hypertension	8.62	7.72	7.87	3.81	8.87
Depression	20.60	10.00	11.63	9.97	11.57
Depression screen	27.47	33.28	17.70	10.45	28.67
Electrocardiogram	8.16	23.53	21.92	18.74	23.11
HbA1c	33.39	22.26	31.11	33.02	38.24

Table 6: Relative absolute errors in prevalence between t_{initial} and t_{final} at the national level for each racial group and health outcome, focused on health outcomes for a cohort in 2018 Q1. Values provided are percentage points.

I Exacerbation, Minimization, and Sign Switches in Disparity Assessment Error

We also consider whether delays consistently lead to over- or under-estimates of the true disparity in absolute terms or whether the direction of the disparity changes entirely. We call the case of an over-estimate an *exacerbation* and the case of an under-estimate a *minimization*. When we observe exacerbation, the disparity at t_{initial} appears higher than it actually is at t_{final} (and in minimization, lower than it actually is). Exacerbations may be a concern when there are limited resources, but the consequences for not intervening are small. Minimizations may be a concern when there

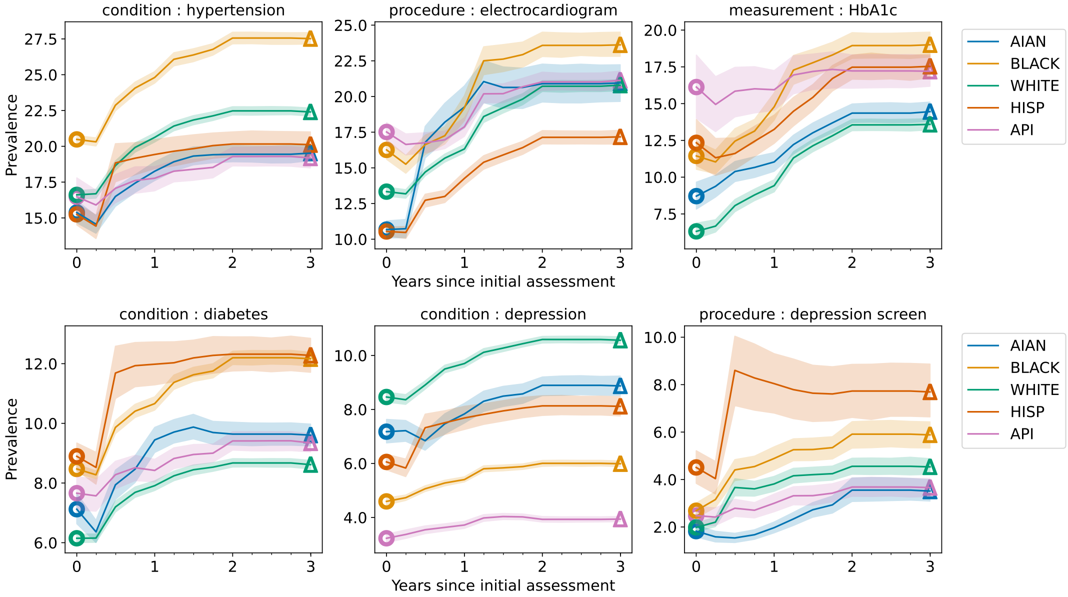

Figure 9: Simulations at the national level for a fixed cohort of patients from 2017 Q1.

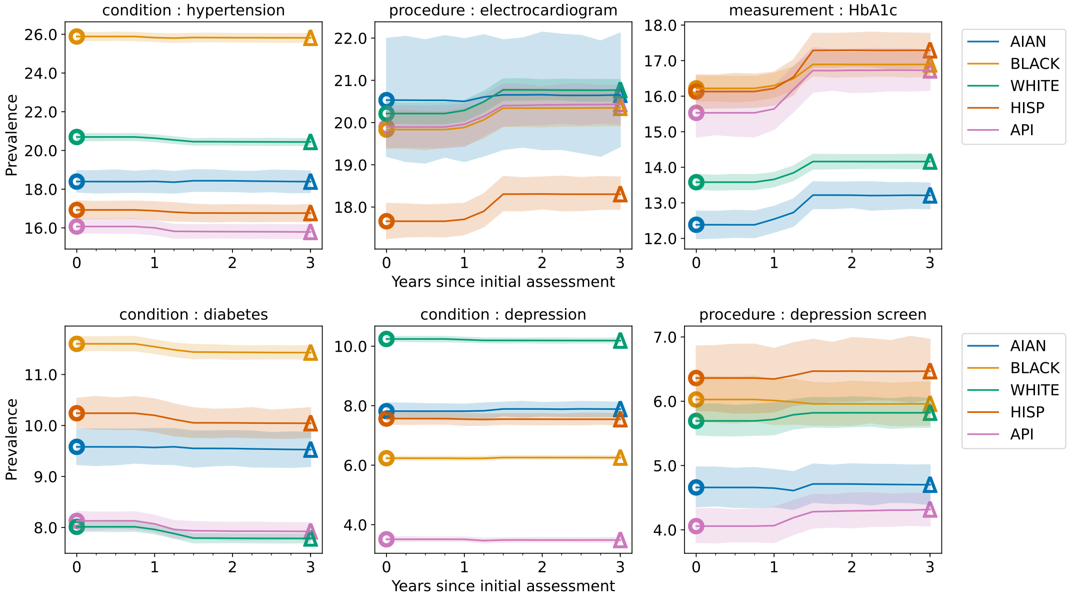

Figure 10: Simulations at the national level for a fixed cohort of patients from 2019 Q1.

are serious health concerns that would require immediate intervention. *Sign changes* (*i.e.*, the direction of the disparity changes, which can be either exacerbations or minimizations) are a problem in either setting as they would distort any meaningful takeaways related to the disparity.

Table 7 summarizes these trends by comparing *disparity estimates* (*i.e.*, pairwise differences in prevalence across groups, for each outcome) between t_{initial} and t_{final}, at the national level. Delays are more likely to minimize the true disparity for three outcomes; *i.e.*, premature evaluations often underestimate disparities. But delays are more likely to exacerbate the true disparity for the other three outcomes; *i.e.*, premature evaluations also often overestimate disparities. Perhaps most concerning is the high rate of sign switching for an outcome like electrocardiogram procedures; *i.e.*, premature evaluations can be wrong about the direction of disparity on average 14% of the time.

J BIFSG: Impact on Error Metrics

BIFSG improves average prevalence error (Figure 12), but the direction of error changes in some cases (Figures 5 and 14). We also visualize worst-case prevalence error, or the highest absolute gap

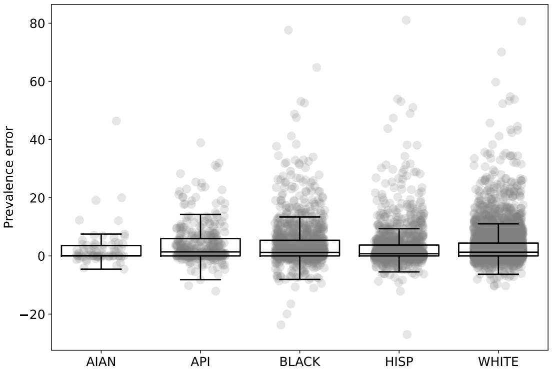

Figure 11: Distribution of prevalence errors for all conditions averaged across all bootstrapped samples at the practice level. We remove all practice-level prevalence estimates that involve fewer than 10 patients on average to preserve patient privacy.

	%Exacerbation	%Minimization	%Sign Switch
Diabetes	51.6	41.2	7.2
Depression	21.2	73.8	5.0
Hypertension	29.4	64.6	6.0
Depression screen	56.4	32.8	10.8
Electrocardiogram	40.0	46.0	14.0
HbA1c	43.4	42.8	13.8

Table 7: Comparison of errors in disparity estimation. We classify sign switches (*i.e.*, rank order changes for the pairwise comparisons, which can be either exacerbations or minimizations as well) first, and then classify the remaining errors as exacerbations or minimizations.

in prevalence error across groups. Furthermore, BIFSG mitigates disparity error in three out of six outcomes (see Figure 13).

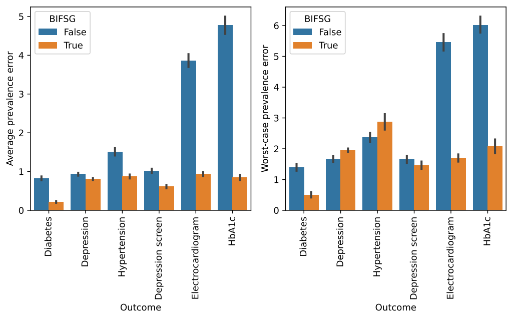

Figure 12: BIFSG reduces the average prevalence error for all outcomes and the worst-case prevalence error for most outcomes.

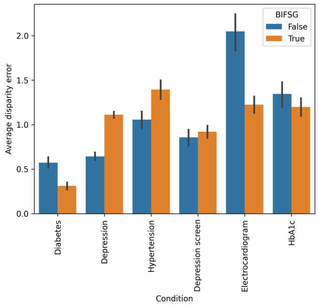

Figure 13: BIFSG mitigates error in disparity assessment (disparity error) in three out of six outcomes, but not the others.

K Delayed Reporting across Consecutive Disparity Assessments

The impact of delayed reporting on disparity assessments cannot be disentangled from the real-world setting in which disparity assessments may be used — *e.g.*, to inform state and local healthcare organizations about any serious health equity gaps and to produce timely interventions. In a real-world setting, monitoring would be conducted at regular intervals for different cohorts (*e.g.*, 2018 Q1, 2018 Q2, 2018 Q3, etc.). As a result, we now let the clock run beyond a single quarter in 2018 and examine delays in real time. Figure 15 illustrates a full year of disparity assessments, where the evaluation of t_{initial} (*i.e.*, 2018 Q1 to Q4) and the improved evaluation of t_{3Yrs} appear together. Note that at each time step, we only present estimates from t_{initial} and t_{3Yrs}. For each outcome we focus on a single pairwise comparison between a White and non-White group, whichever group experiences the largest average disparity at t_{3Yrs} for that outcome over the course of 2018.

This figure underscores that delay can manifest in *every* time step — a formidable challenge for continuously monitoring disparities. In principle, evidence of disparities would trigger interventions as soon as possible and regular monitoring would subsequently reveal improvements over time. But as long as delays continue to distort ground truth disparity estimates at the same pace as assessments are conducted, decision-makers may need to choose between estimating disparities inaccurately or monitoring health disparities at a slower pace.

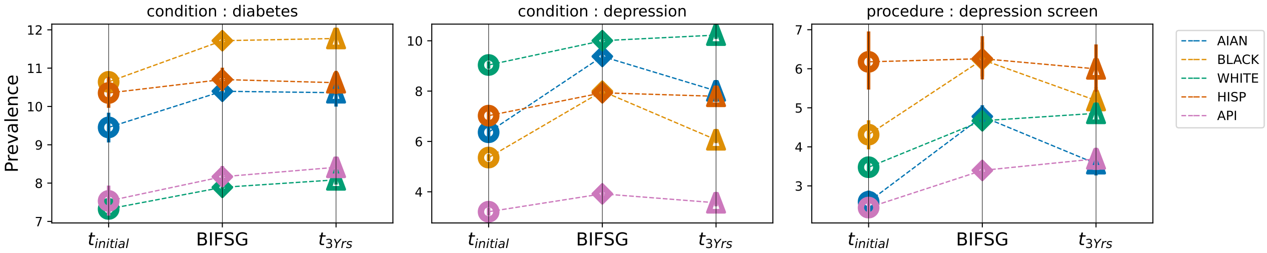

Figure 14: BIFSG over-estimates prevalence for several minority race groups (*e.g.*, Black and AIAN patients) across several outcomes, though average prevalence error is improved. In each subplot, the y-axis denotes the estimated prevalence. Values for $t_{initial}$ and t_{3Yrs} match the same national values as shown in Figure 7.

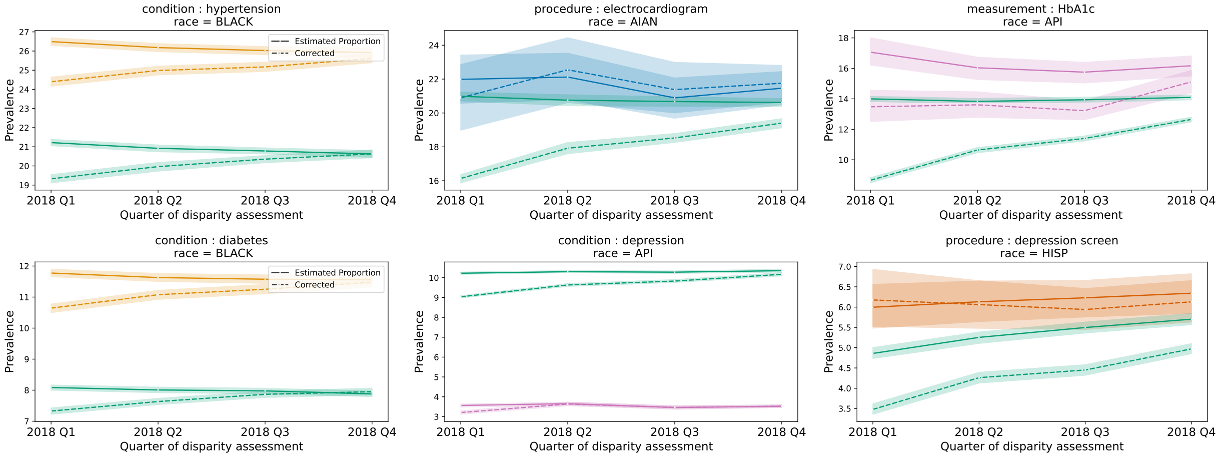

Figure 15: Comparison of prevalence estimates for different health outcomes at quarterly intervals starting in 2018. This figure invokes a real-world example of conducting regular disparity estimates given incomplete or delayed information. For each health outcome, we produce t_{initial} estimates (solid line) and t_{3Yrs} estimates (dashed line) for patients in a minority group (orange, yellow, blue, or purple), in comparison to estimates for White patients (green). We select pairwise comparisons for whichever groups experience the largest average disparity at t_{3Yrs}.