Evidence generation, decision making, and consequent growth in health disparities

Anirban Basu[a,b,1] and Kritee Gujral[a]

[a]The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, University of Washington, Seattle, WA 98195; and [b]Health Care Program, The National Bureau of Economic Research, Cambridge, MA 02138

Edited by Charles F. Manski, Northwestern University, Evanston, IL, and approved April 30, 2020 (received for review November 15, 2019)

Evidence is valuable because it informs decisions to produce better outcomes. However, the same evidence that is complete for some individuals or groups may be incomplete for others, leading to inefficiencies in decision making and growth in disparities in outcomes. Specifically, the presence of treatment effect heterogeneity across some measure of baseline risk, and noisy information about such heterogeneity, can induce self-selection into randomized clinical trials (RCTs) by patients with distributions of baseline risk different from that of the target population. Consequently, average results from RCTs can disproportionately affect the treatment choices of patients with different baseline risks. Using economic models for these sequential processes of RCT enrollment, information generation, and the resulting treatment choice decisions, we show that the dynamic consequences of such information flow and behaviors may lead to growth in disparities in health outcomes across racial and ethnic categories. These disparities arise due to either the differential distribution of risk across those categories at the time RCT results are reported or the different rate of change of baseline risk over time across race and ethnicity, even though the distribution of risk within the RCT matched that of the target population when the RCT was conducted. We provide evidence on how these phenomena may have contributed to the growth in racial disparity in diabetes incidence.

evidence-based medicine | health disparity | diabetes incidence | treatment effect heterogeneity

Evidence on the comparative effectiveness of alternative treatments is usually generated using randomized clinical trials (RCTs). These trials are conducted across multiple settings, where patients are enrolled by physicians who are the principal investigators of the studies. In some cases, there are multiple trials that either are funded by different organizations or occur at different points in time, with a different set of investigators across these trials. The results from these multiple trials may oppose or reinforce each other's findings. In the early 1990s, the framework of evidence-based medicine (EBM) was developed to summarize the body of evidence on a research question and to use that summary to be the current best evidence from clinical care research in the management of individual patients. The fact that average results from EBM do not apply to individual patients is well known (1). However, EBM seems to suggest that after careful curation based on exclusion and inclusion criteria from multiple studies, if all patients in the target population were to be treated with a given treatment, the realized population-level outcomes would reflect those obtained in the EBM analysis for that treatment (1–3). Many clinical societies, therefore, swear by EBM for recommending the standard of care, for developing clinical guidelines, and for issuing directives for clinical care in the population. Nevertheless, in practice, information about clinical care for individual patients may additionally arise from a variety of settings, e.g., previous smaller studies that have focused on specific groups of patients, learning by doing in clinical practice, and social interactions. To some extent EBM results may directly conflict with information acquired through these different channels. In the presence of treatment effect heterogeneity, these sequential processes of RCT enrollment with prior information, information generation, and dissemination via EBM and consequent treatment choice decisions have not been jointly examined formally.

We show that, in the presence of treatment effect heterogeneity over some baseline measures of patient risk, the currently accepted process of evidence generation and application may have an unintended consequence of increasing disparities in health outcomes across different subgroups, including racial and ethnic categories (RECs), even when there are no disparities in treatment choices conditional on the baseline risk. Disparities across RECs are the main focus for this paper, but the underlying phenomena studied could be applied to study disparities generally across other subgroups.

Specifically, we develop a theoretical model to study the behavior of enrollment in RCTs, the evidence-based recommendation from such RCTs, and the consequent impact on treatment selection in practice. An overview of this information flow and behaviors can be found in Fig. 1. Throughout this framework, we consider the individual decision maker to be the patient–physician dyad, referred to as the patient. Individual patients in a target population start with some beliefs about the incremental benefits of the treatment and consider whether to enroll in an RCT designed to test a treatment compared to a control or to use the treatment without enrolling in an RCT, should that treatment be available outside of the RCT (Fig. 1). We specifically focus on treatments that are already available in the population and not on innovations being evaluated by a regulatory agency such as the US Food and Drug Administration

Significance

When treatment effects are heterogeneous, reliance on average effects from clinical trials also means relying on the distribution of baseline risk in the trial. Even when trials are representative of the then population, a drift in the risk distribution over time, especially if differential across race or SES categories, could generate disparities in outcomes when decision making on treatment choices continues to rely on the average effect from the trial. Based on our theoretical model, we illustrate that unintended phenomena involving production and use of average evidence from large clinical trials could explain up to 10.5% of the recent growth in racial disparity in diabetes incidence in the United States.

Author contributions: A.B. designed research; A.B. and K.G. performed research; A.B. and K.G. analyzed data; and A.B. and K.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: The analytical Stata dataset and a variable dictionary are available at the Open Science Framework: https://osf.io/gr29s/files/.

[1] To whom correspondence may be addressed. Email: basua@uw.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920197117/-/DCSupplemental.

First published June 8, 2020.

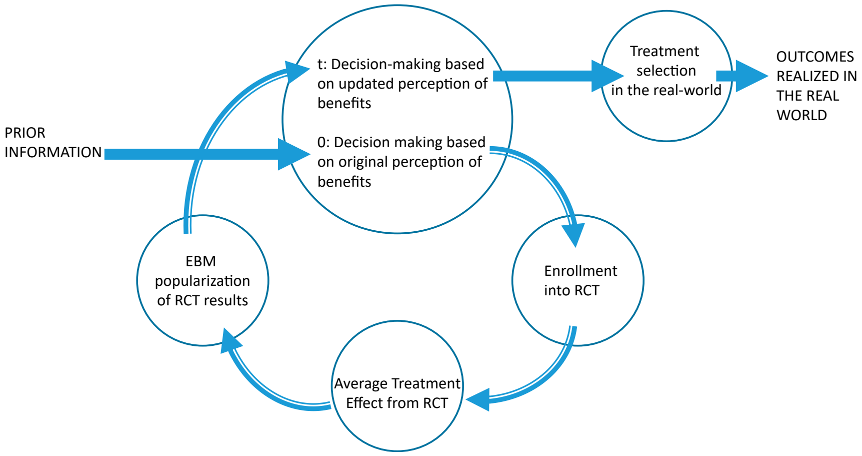

t: Decision-making based on updated perception of benefits
0: Decision making based on original perception of benefits
PRIOR INFORMATION
Treatment selection in the real-world
OUTCOMES REALIZED IN THE REAL WORLD
Enrollment into RCT
Average Treatment Effect from RCT
EBM popularization of RCT results

Fig. 1. Flow of information and behaviors in clinical evidence generation and practice.

(FDA). Similar models have been used to study enrollment in FDA trials by Philipson (4) and Malani (5).[†] Depending on those who enroll, an average effect is estimated from the RCT. This effect determines the evidence-based recommendation, which, when popularized through EBM, updates the beliefs of all individual patients in the population for future years. Finally, at any given point, individual beliefs and the costs of acquiring treatment drive decisions about getting the treatment in practice. The realized outcomes in the population are based on these individual treatment choice decisions, shaped by EBM recommendations.

By formalizing these processes of information flow and behavior, we show that at every step where new information affects behavior, whether in terms of RCT enrollment or treatment selection in practice, there is a potential for the unintended consequence of inducing disparities in outcomes across patients in different RECs due to both the differences in the current distribution of baseline risk across RECs and the differences in the rate of change of baseline risk across RECs over time.

We apply this model to a case of diabetes prevention to illustrate how these phenomena may have contributed to the growth of racial disparity in diabetes incidence. Specifically, our model predicts that due to the popularization of the intensive lifestyle modification intervention for diabetes prevention from average results of RCTs, at-risk individuals who fail to engage in lifestyle modification due to its shadow costs do not take up the alternative metformin therapy, which is equally effective in high-risk patients. In addition, if the distribution of high risk shifts at a different rate for subgroups, a natural disparity in health outcomes will emerge across those subgroups. We show that treatment choice behaviors in diabetes prevention are in line with our theoretical model's predictions. Moreover, counterfactual predictions from our model suggest that shortcomings of the current evidence generation and its application could explain up to 10.5% of the growth in racial disparity in diabetes incidence. We conclude by discussing ways in which such shortcomings of EBM and the RCT infrastructure, in general, can be alleviated.

In the next sections, we develop the theoretical model for the abovementioned sequence of behaviors and illustrate this framework using an application of EBM for diabetes prevention.

A Motivating Example: Diabetes Incidence

Type 2 diabetes is one of the most prevalent chronic diseases of our time. The Centers for Disease Control and Prevention (CDC) estimates that 1 in 10 Americans have diabetes and 1 in 3 have prediabetes, i.e., risk of developing diabetes. However, a majority of prediabetics are unaware of their prediabetic status. Based on over two decades of research, the CDC-led National Diabetes Prevention Program (NDPP), which is a partnership of public and private organizations working to prevent or delay type 2 diabetes, helps individuals make lifestyle changes to prevent or delay type 2 diabetes and other serious health problems. A key component of the NDPP is a lifestyle change program that helps people lose 5 to 7% of their body weight through healthier eating and 150 min of physical activity of moderate intensity per week, such as brisk walking. This program and its guidelines are rooted in results from rigorously conducted randomized clinical trials, published in 2002, demonstrating that people with prediabetes who take part in a structured lifestyle change program could cut their risk of developing type 2 diabetes by 58% (71% for people over 60 y old) (6).

Diabetes incidence had been rising (+4.7% annual percentage change, P value < 0.001) in the United States over last two decades until 2008 (7). Since 2008, there has been a decrease in such incidence (−5.4% annual percentage change, P value = 0.09) (7). Efforts continue to reduce type 2 diabetes through the NDDP lifestyle change intervention, which includes targeted screening, as well as population approaches to improve healthy food availability, diabetes awareness, and education and walkability of communities (8).

However, while the trends in diabetes incidence for non-Hispanic Whites were roughly similar to the overall trends showing a decline in incidence since 2008, there were no such declines for both non-Hispanic Blacks and Hispanics. The increase in diabetes incidence continued unabated among these two subgroups through 2012, thereby widening the health disparities from about 3/1,000 pre-2008 to 7/1,000 by 2012 (Fig. 2).

There can be competing explanations for this increase in health disparity. It is possible that the DPP program was not effective in these two subgroups and/or, even if the program was effective, its uptake remained compromised in these subgroups, compared to Whites. It is also possible that the EBM approach that led to the guided program misapplied information from the RCTs, resulting in suboptimal outcomes in some subgroups compared to others. Exploring this latter channel is the focus of this paper.

[†] Both of their works use Roy's (9) model but in cases of new technology assessments where the treatment is not available outside the RCT and likely there is no anticipation of heterogeneity in the absence of prior information. That makes their models special cases of the one presented here.

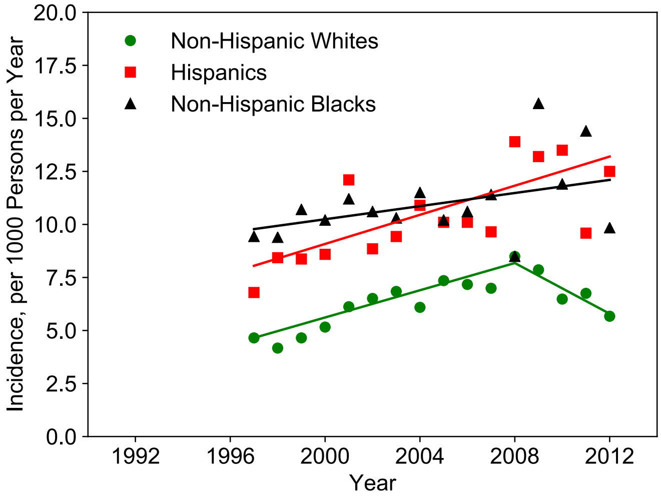


Fig. 2. Diabetes incidence 1997 to 2012 by race. Data from ref. 7.

A Theoretical Model for Evidence Generation and Disparities

Consider θ to be a scalar risk, representing a set of factors over which the "true" treatment effects (e) vary; i.e., $e_i = e(\theta_i)$ represents the conditional benefits function and i represents an individual. Note that here, θ is a heterogeneity parameter, which has a distribution in the population, although for an individual patient, it has a deterministic value (similar to age). Without loss of generality, let $\theta \in \Re$, $E(\theta) = 0$, and $e(\theta) \sim \mathcal{N}(\mu_e(\theta), \sigma_e^2)$, where

$$\mu_e(\theta) = \alpha_1 + \alpha_2\theta, \qquad [1]$$

and $\alpha_1, \alpha_2 \in \Re$. Therefore, the true population average treatment effect parameter is α_1.

Let the perception about the incremental benefits of treatment be given by $r(\theta)$, where

$$r(\theta_i) = e(\theta_i) + \epsilon_i, \qquad [2]$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_r^2)$. This signifies that although, on average, perceptions align with true benefits, there could be substantial noise for any given individual. One may allow σ_r^2 to vary with θ. However, there is no expected direction for $\frac{\partial \sigma_r^2}{\partial \theta}$.[‡] Since including such nuance can distract from the main message that even under basic circumstances, enrollment in RCT can be systematic, we do not consider heteroscedasticity for the sake of simplicity. Therefore, Eq. **2** represents a classical measurement error problem in the perception of benefits from treatment, which can form from previous studies, from learning by doing in practice, and also from social interactions. The implications for relaxing the dependence between $r(\theta)$ and $e(\theta)$ are considered below.

At this point, let patients face a choice of enrolling in an RCT or seeking treatment outside of an RCT (denoted by "OUT"). Let the costs (incremental to the corresponding control intervention) of obtaining the treatment be given as $C_{ij} = c_j + m_j(\theta_i)$, where j = RCT or OUT. c_j represents demand prices for treatment and m_j represents the shadow costs of getting access to the treatment, e.g., travel time, treatment time, etc.[§] For the sake of simplicity, cost components are not considered to be stochastic. Shadow costs are assumed to vary with θ.

Next, consider three behaviors corresponding to one cycle of information flow, shown in Fig. 1, which would be represented using a two-time-period model. In the first period, patients have the choice to enroll in an RCT, which produces results that are then applied as treatment guidance. In the second period, the patient makes treatment selection based on the patient's updated perceptions of treatment benefits.

Enrollment in an RCT. Following a standard Roy's (9) model on sorting behavior, let S_i, an indicator for an individual enrolling in an RCT, be given by

$$S_i = I(U_i^* \geq 0), \qquad [3]$$

where U_i^* is the patient's latent net utility for enrolling, which is expressed as

$$U_i^* = [\pi_R{}^* r(\theta_i) - C_{i_{\mathrm{RCT}}}(\theta)] - [r(\theta_i) - C_{i_{\mathrm{OUT}}}(\theta)], \qquad [4]$$

and π_R is the known random probability of receiving treatment within the RCT (and $1 - \pi_R$ is the probability of being assigned to the control group). Given that no individual with negative perceived benefits of treatment will enroll in the RCT, the probability π of enrolling in the RCT is given as

$$\pi(\theta) = Pr\left(r(\theta) \leq \frac{C_{\mathrm{OUT}}(\theta) - C_{\mathrm{RCT}}(\theta)}{1 - \pi_R}\right) = \Phi(h(\theta)), \qquad [5]$$

where $\Phi()$is a cumulative standard Gaussian distribution and

$$h(\theta) = \frac{\substack{(c_{\mathrm{OUT}} - c_{\mathrm{RCT}}) + (m_{\mathrm{OUT}}(\theta) - m_{\mathrm{RCT}}(\theta)) \\ -(1-\pi_R)(\alpha_1 + \alpha_2\theta)}}{(1 - \pi_R)\sqrt{\sigma_e^2 + \sigma_r^2}}. \qquad [6]$$

The above implies that only those individuals whose positive perceived benefits are less than the weighted incremental costs of accessing the treatment outside the RCT will enroll in the RCT (10). The probability of enrolling in the RCT will depend on the true benefit parameters, differential costs, and the noise in perceived benefits. More importantly, as θ increases, the probability of enrolling into the RCT will decrease if the difference in the shadow costs of accessing treatment outside versus inside the RCT decreases with θ. This is possible when individuals who would have benefited the most from treatment find it more difficult than others to enroll in the RCT due to access and other issues. The probability will also decrease if the true benefits of treatment increase faster than the difference in costs of accessing treatment outside versus inside the RCT over θ. This is often the case when the treatment in question is already available outside the RCT and also is covered through insurance. Consequently, individuals who anticipate higher benefits, even with the additional noise, are less likely to enroll in the RCT.

It should be noted that even if there was no systematic anticipation of benefits, i.e., Eq. **2** was $r(\theta_i) = \epsilon_i$, and $Corr(r(\theta), e(\theta)) = 0$, enrollment probability in the RCT may still vary over θ, but only through the differential shadow costs of obtaining treatment inside versus outside the RCT.

Finally, we have assumed π_R to be constant and known a priori. Eq. **5** suggests that as long as $\pi_R < 1$, perceived benefits would still influence treatment selection. Some adaptive designs for RCTs do allow π_R to vary over time and risk groups. To what extent such designs can be used to over/undersample and fix distortions in enrollments remains to be studied.

[‡] That is, subgroups at high risk may have smaller or larger variation in beliefs. If heteroscedasticity increases with θ, and so do expected benefits, then the relative direction for enrollment into an RCT, as evident below, remains undetermined. This is because expected higher benefits make patients less likely to enroll, while higher uncertainty makes patients more likely to enroll.

[§] It is assumed that the cost of accessing the placebo or the control group within or outside the RCT is the same.

Average Treatment Effect Parameter in an RCT and Its Popularization through EBM. While the population average treatment effect (ATE) parameter, following Eq. **1**, is given as

$$\mathrm{ATE_{POP}} = \int e(\theta)dF(\theta) = E(e(\theta)) = \alpha_1 + \alpha_2 E(\theta) = \alpha_1, \quad \textbf{[7]}$$

the average treatment effect parameter for the RCT is given by

$$\begin{aligned}\mathrm{ATE_{RCT}} &= \int_0^1 \frac{\pi(\theta)e(\theta)}{\bar{\pi}(\theta)}dF(\theta) \\ &= \alpha_1 + \frac{COV(\pi(\theta), e(\theta))}{\bar{\pi}(\theta)},\end{aligned} \quad \textbf{[8]}$$

where $\bar{\pi}(\theta) = E(\pi(\theta)) = E(\Phi(h(\theta)))$. In other words, the enrollment probabilities will weight the conditional benefits function to define the target parameter for the RCT. When there is no correlation between the probability of enrollment and the true benefits ($COV(\pi(\theta), e(\theta)) = 0$), $\mathrm{ATE_{RCT}} = \mathrm{ATE_{POP}}$. In cases where individuals who would truly realize higher benefits of treatment are less likely to enroll in the trial, i.e., $COV(\pi(\theta), e(\theta)) < 0$, $\mathrm{ATE_{RCT}} < \mathrm{ATE_{POP}}$ and vice versa (10).

One should note that the average effect conditional on θ within an RCT, even with nonrandom selection, is a consistent estimator of the true conditional population average effect; i.e., $E(\mathrm{ATE_{RCT}}(\theta)) = e(\theta)$. Once these conditional effects are estimated, they can be reweighted using the population distribution of θ to recover mean population-level treatment effect parameters. However, most, if not all, trials do not attempt to establish such nuanced risk-based distribution of treatment effects. Some large RCTs carry out subgroup analyses to assess the heterogeneity of treatment effects. It is clear from Eq. **8** that any average estimator over any restricted support of the distribution of θ will have the same issues as $\mathrm{ATE_{RCT}}$.

Let the estimated effect from the RCT be given as $\bar{y} \sim \mathcal{N}(\mathrm{ATE_{RCT}}, s_{\bar{y}}^2)$. EBM would popularize this estimated average effect from the RCT, which would impact the beliefs of individuals in the population. Denoting original (first-period) and updated (second-period) beliefs with subscripts "0" and "*t*," respectively, updated beliefs following EBM recommendations, assuming a standard Bayesian updating with Gaussian conjugates, are given as

$$r_t(\theta) = wr_0(\theta) + (1 - w)\bar{y}, \quad \textbf{[9]}$$

where $r_t(\theta) \sim \mathcal{N}(\mu_{r_t}(\theta), \sigma_{r_t}^2)$ and

$$\mu_{r_t}(\theta) = w(\alpha_1 + \alpha_2\theta) + (1 - w)\mathrm{ATE_{RCT}} \quad \textbf{[10]}$$

and

$$\sigma_{r_t}^2 = w\sigma_{r_0}^2, \quad \textbf{[11]}$$

where $w = s_{\bar{y}}^2/(s_{\bar{y}}^2 + n\sigma_{r_0}^2)$ is the evidence weight of prior beliefs compared to RCT, with n being the sample size of the RCT. Even when the RCT provides a consistent estimator for the population-level average treatment effect but provides no evidence on heterogeneity, the application of the average RCT result through EBM has an attenuation effect w on the perceived treatment benefit heterogeneity over θ.

Implications for Disparities Based on Updated Beliefs, Treatment Selection, and Realized Outcomes. A comparison of how realized outcomes in the population, resulting from clinical practice, would differ from those in an idealized world can be made by comparing treatment choices patients make based on updated beliefs and the variance in their beliefs, given in Eq. **10**, with those they would have made if $e(\theta)$ were known. A formal representation of the discussion that follows is provided in *SI Appendix*. In this section, we provide an intuition for how these differences between the idealized world and the case of clinical practice may generate disparities in health outcomes across subgroups.

In the idealized world, the probability of treatment choice, i.e., $Pr((e_i(\theta) - C_{i\mathrm{OUT}}(\theta)) > 0)$, increases with θ as long as the costs of accessing the treatment are not rising at a faster rate than the benefits with respect to θ (i.e., $\alpha_2 > \frac{\partial m_{\mathrm{OUT}}(\theta)}{\partial\theta}$). Treatment selection in clinical practice, however, will be driven by the updated beliefs about the effect of treatment following EBM. In this setting, the probability of treatment choice will be given by $Pr((r_t(\theta_i) - C_{i\mathrm{OUT}}(\theta)) > 0)$. Compared to the ideal setting, since $w < 1$ and the perceived benefits get shrunken toward the $\mathrm{ATE_{RCT}}$, the probability of treatment selection in practice will diverge from the ideal setting and more so at higher values of θ. This holds true with or without self-selection into the RCT. With self-selection, especially when individuals who would truly realize benefits of treatment are less likely to enroll in the trial ($COV(\pi(\theta), e(\theta)) < 0$), the sensitivity (of the divergence of treatment selection in practice vs. ideal) to θ becomes amplified. This is not to say that the RCT was not valuable since we do not know the nature of treatment selection pre-RCT. Nonetheless, it shows how the average effect from RCTs accentuates the decision errors in practice compared to the ideal setting, especially for individuals with high values of θ.

It is important to note here that, conditional on θ, our model has no implications for disparities in treatment choices. It is true that the shadow price of obtaining care may differ across subgroups, independent of θ. However, that is not the main premise of our analysis. In our model, even when the shadow prices are the same across subgroups, the differences in the consequent realized outcomes between the idealized and clinical practice settings would depend on θ because of the universal inefficiency of treatment choices at specific θ levels induced by the average results. The direction or nature of this dependence would vary based on the relative magnitudes of the $Corr(e(\theta), r_t(\theta))$, w, and on the sign of α_2. The key point here is that the difference between realized benefits and ideal benefits decreases (or increases) with θ and could become negative (or more positive) at higher values of θ. This intuition captures the main implication of our current evidence production and application infrastructure for disparities. It says that any two subgroups of patients with a different distribution of θ will, on average, experience different levels of realized outcomes compared to ideal outcomes, potentially giving rise to the growth in disparities in outcomes. These implications hold even when the RCT was representative of the contemporaneous distribution of θ in the population (i.e., $E(\bar{y}) = \alpha_1)$) or when patients did not anticipate the directions of treatment effect heterogeneity over θ (i.e., $\rho = 0$). Differences in the distribution of θ and also a differential shift in those distributions over time across subgroups will still induce disparities in outcomes.

In the next section, we provide empirical evidence illustrating how these mechanisms may have given rise to observed racial disparity in outcomes.

Evidence-Based Medicine on Diabetes Incidence and Disparities

As mentioned earlier, the NDPP and its specific guidelines, including the key lifestyle change program, stem from rigorously conducted randomized clinical trials, published in 2002, which we now discuss in the context of the EBM approach.

Diabetes Prevention Program Trial, Results, Guidelines, and Treatment Effect Heterogeneity. The Diabetes Prevention Program (DPP) was a three-arm randomized clinical trial testing strategies for preventing or delaying the development of type 2 diabetes in high-risk individuals with elevated fasting plasma glucose (FPG) concentrations and impaired glucose tolerance (IGT). Patients were randomly assigned to one of three intervention groups: 1) standard lifestyle recommendations plus metformin (dose of 850 mg twice daily), 2) standard lifestyle recommendations plus placebo (twice daily), or 3) an intensive program of lifestyle modification focusing on a healthy diet and exercise.

The DPP trial aimed to recruit a large and diverse cohort of individuals at high risk for developing type 2 diabetes. It focused on recruiting obese individuals from a wide range of age groups and racial backgrounds. The DPP included 27 clinical centers in the United States that began the recruitment process in June 1996 with a randomization goal of 3,000 participants (12). To a large extent, the DPP remains a gold standard trial among large RCTs in the United States.

The DPP trial results were published in 2002, demonstrating a reduction in diabetes incidence of 58% using the lifestyle intervention and a reduction of 31% using metformin, compared to placebo (6). These effects were similar across gender and across racial/ethnic backgrounds. Given that the DPP trial was conducted with a diverse and multiracial cohort, the study emphasized the applicability of their findings "to the ethnically and culturally diverse population of the United States" (ref. 6, p. 6). In 2003, the United States Preventive Services Task Force (13) assigned intensive lifestyle weight-loss interventions a B grade based on a review of the literature, with fair to good evidence for modest, sustained weight loss. Specifically, the DPP clinical guidelines stated "lifestyle modification was nearly twice as effective (compared to metformin) in preventing diabetes (58 vs. 31% relative reductions, respectively)" (ref. 14, p. S65). Clinical guidelines also highlighted some subgroup-level heterogeneity, stating "[i]n the DPP, metformin was about half as effective as diet and exercise in delaying the onset of diabetes overall, but it was nearly ineffective in older individuals (≥ 60 y of age) or in those who were less overweight ($\mathrm{BMI} < 30\mathrm{kg/m^2}$)" (ref. 14, p. S66). Conversely, metformin was as effective as lifestyle modification in individuals age 24 to 44 y or in those with a body mass index (BMI) $\geq 35\ \mathrm{kg/m^2}$.

In the interest of pursuing more efficient, effective, and patient-centered outcomes, Sussman et al. (11) analyzed 95% of the DPP trial data to test whether participants in the DPP trial varied in their likelihood of receiving benefits from metformin or the lifestyle intervention. They divided up the trial populations into quarters, depending on patients' preintervention risk levels. Their analysis showed that patients with a high risk of diabetes varied substantially in their likelihood of benefiting from the DPP's treatments, depending on their levels of baseline risk (Fig. 3). The lifestyle intervention was six times more effective in reducing absolute risk for patients in the highest-risk quarter, compared to those in the lowest-risk quarter, while still being beneficial for patients in the lowest-risk quarter. However, the beneficial effects of metformin were found to be concentrated entirely in the high-risk quarter (an enormous effect of 25% reduction in absolute risk, which was not statistically different from the effect of lifestyle intervention), but there were no benefits from metformin use in the lowest-risk quarter.

Enrollment in the DPP Trial. We first examine the generalizability of the DPP trial to the target population using Sussman et al.'s (11) construct of patients' preintervention risk levels. To identify the prediabetic US population, i.e., the population at risk for progression to diabetes, we use nationally representative data from the National Health and Nutrition Examination Survey (NHANES) 2005 to 2012 (i.e., four cross-sectional and consecutive 2-y survey cycles). Since IGT and FPG levels were primary risk factors for identifying the target population and these measurements were available only starting in 2005, we are unable to examine data prior to 2005. For the DPP trial data used in our analysis, we rely on estimates provided in Kent et al. (15) and Sussman et al. (11).

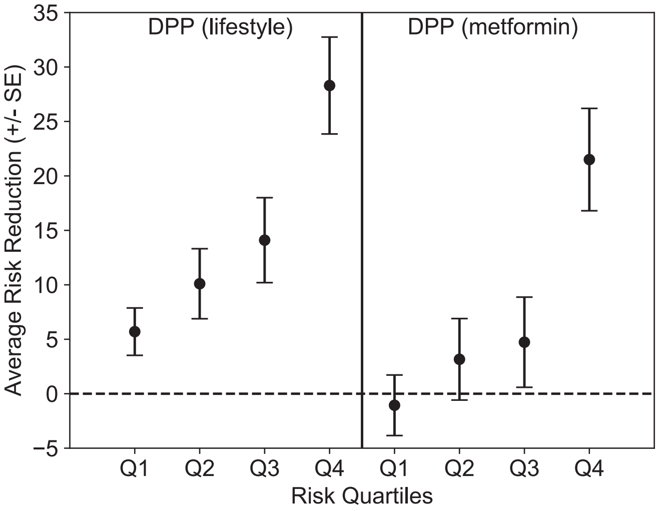


Fig. 3. Reanalysis of treatment effect heterogeneity in the DPP RCT with respect to continuous baseline risk. Data from ref. 11.

We construct baseline risk distributions for the at-risk population by using the inclusion and exclusion criteria applied in the DPP study (16). We include patients comparable to those in the DPP trial, applying the same age and BMI criteria. We consider patients to be prediabetic if they had IGT, elevated FPG, or self-reported prediabetes. Details of the inclusion and exclusion criteria applied are presented in *SI Appendix*, Tables S1 and S2. The application of these criteria results in an at-risk population consisting of the at-risk population we analyzed consists of 2,108 people, which is representative of 22,540,205 nationally (i.e., 7.5% of the US population, 2005 to 2012).

The population risk distributions by race for each 2-y survey cycle over the quartiles of the risk distribution of the DPP trial are presented in Fig. 4. A formal test for the equality of the risk distribution by race over these four categories was conducted using Fisher's exact test, which shows significant differences for each 2-y cycle. Next, note that during 2005 to 2006, the population risk distribution for each race seems to be similar to that of the DPP. That is, within each DPP risk quartile (which by definition includes 25% of the DPP trial participants), the proportion of population within each racial category is also 0.25. Note further that such relative similarity in risk distribution across race changes after the 2005 to 2006 period. Racial minority groups are more likely to be in the highest DPP risk category each year, compared to non-Hispanic Whites. For example, by 2007 to 2008, over 40% of Hispanics are in the highest-risk category compared to about 25% for both non-Hispanic Whites and Blacks. By 2011 to 2012, over 40% of non-Hispanic Blacks and over 35% of Hispanics were in the highest-risk category, compared to only 30% of non-Hispanic Whites. It is evident that over time, non-Hispanic Blacks and Hispanics have a higher concentration of prediabetes patients in the high-risk quartile, where the effectiveness of lifestyle over metformin appears lowest (Fig. 3). Based on insights developed in *Enrollment in an RTC* and *Average Treatment Effect Parameter in an RCT and Its Popularization through EBM*, over time, the ATE from the DPP trial does not accurately reflect ATE for non-Hispanic Blacks and Hispanics.

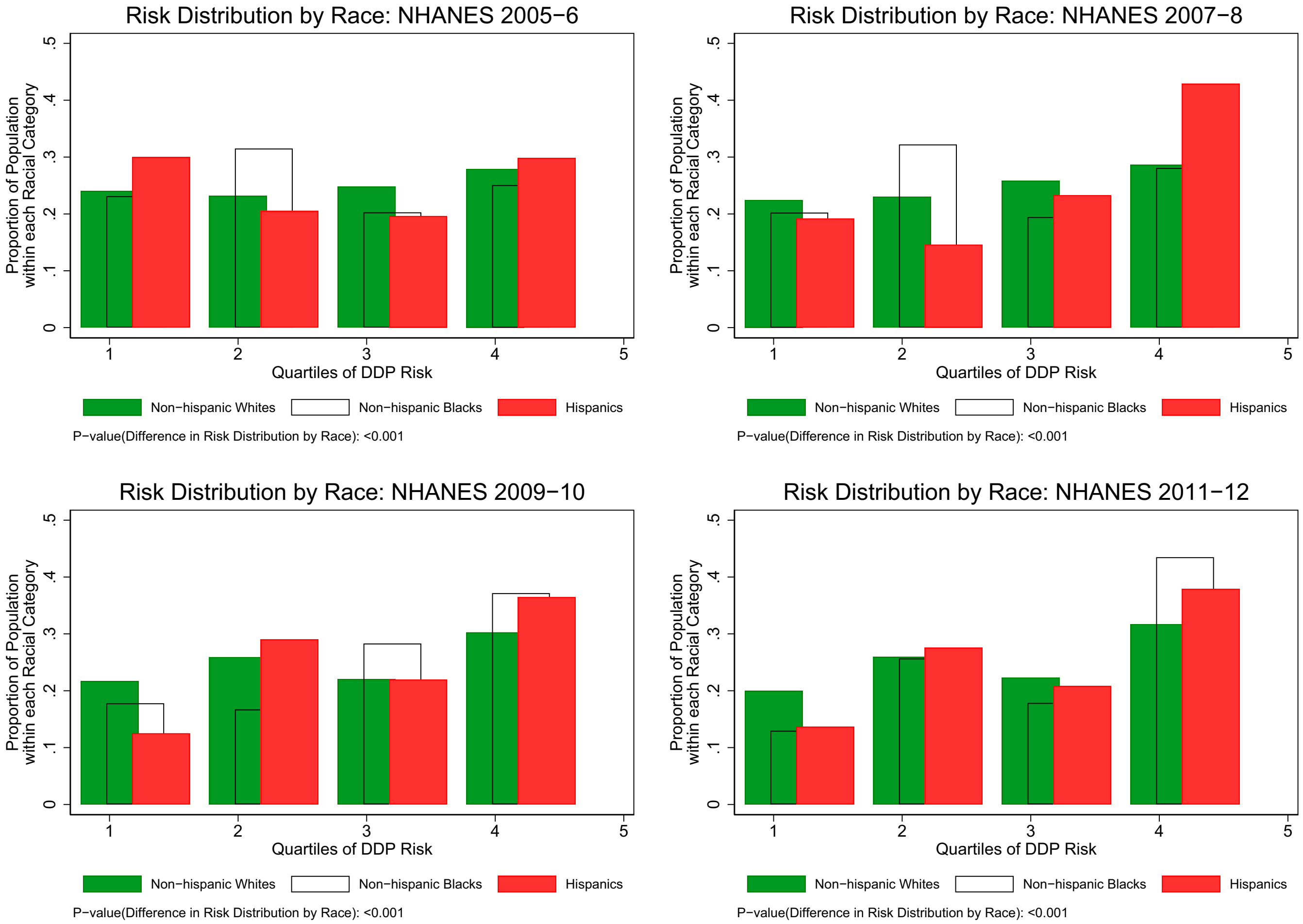


Fig. 4. Baseline risk for diabetes in the US target population (NHANES 2005 to 2012) across quartiles of risk score in the DPP trial.

Implications of Theoretical Models for Treatment Uptake and Disparity in This Example. There are two sets of treatment comparisons in the DPP: 1) lifestyle modification (treatment) versus metformin (control) and 2) metformin (treatment) versus placebo (control). θ represents the baseline risk of developing diabetes. Since the risk distributions in the DPP and in the 2005 target population looked similar, we assume the following: 1) Before the DPP, patients did not have any anticipation of the directions of treatment effect heterogeneity over θ for any of the comparisons (i.e., $Corr(r_0(\theta), e(\theta)) = 0$ for either comparison). 2) The demand price and shadow prices of treatments were the same inside and outside the DPP (i.e., $c_{\mathrm{OUT}}(\theta) = c_{\mathrm{DPP}}(\theta)$ and $m_{\mathrm{OUT}}(\theta) = m_{\mathrm{DPP}}(\theta)$ for both lifestyle intervention and metformin).

These two assumptions taken together imply, according to Eq. **5**, that $\pi(\theta) = \pi \ \forall\theta$. Consequently, $COV(\pi(\theta), e(\theta)) = 0$, and according to Eq. **8**, $\mathrm{ATE}_{\mathrm{RCT}} = \alpha_1$. Therefore, the average effect from the DPP, for either comparison, will be a consistent estimator of the true population average treatment effect. This result would also imply, according to Eq. **9**, that $Corr(e(\theta), r_t(\theta)) = \rho = 0$.

In addition, we assume that the incremental shadow costs of treatment versus control in either comparison are positive, and these incremental shadow costs do not decrease with respect to θ. (i.e., $\frac{\partial m_{\mathrm{OUT}}(\theta)}{\partial\theta} \geq 0$). This is reasonable as it is generally expected that patients who are at higher risk of developing diabetes, due to social, environmental, and medical reasons, are also more likely to face steeper shadow costs of spending more time exercising while facing the same costs for generic metformin. Finally, treatment effect heterogeneity results from the DPP (Fig. 3) suggest the following: 1) The effect of lifestyle modification over metformin is positive on average (i.e., $\alpha_1 > 0$) but decreases over θ (i.e., $\alpha_2 < 0$). 2) The effect of metformin over placebo is also positive on average (i.e., $\alpha_1 > 0$) but increases over θ (i.e., $\alpha_2 > 0$).

Based on the directions of α_1 and α_2 in each comparison, we would expect that, in the ideal setting, the probability of engaging in lifestyle intervention should decrease over θ, since metformin is an equally effective treatment at higher values of θ with lower shadow costs. Similarly, the probability of using metformin should commensurately increase over θ.

However, in practice, with the DPP average results, perceived benefits will be

$$\mu_{r_t}(\theta) = \alpha_1 + w\alpha_2\theta, \qquad [12]$$

which, when compared to perceived benefits under the idealized environment, and assuming that the DPP was large enough so that $\sqrt{w\sigma^2_{r_0}} \sim \sigma_e$, would lead us to expect that the probability of engagement in lifestyle intervention would be higher at each level of θ compared to the ideal setting (as $\alpha_2 < 0$), and this difference would increase with θ. In contrast, we would expect that the probability of using metformin would be lower at each level of θ compared to the ideal setting (as $\alpha_2 > 0$), and again this difference would increase with θ. The empirical question that remains is whether or not the inefficiently excess use of lifestyle intervention will be sufficient to offset the inefficiently inadequate use of metformin at higher levels of θ, purely from the perspective of preventing diabetes.

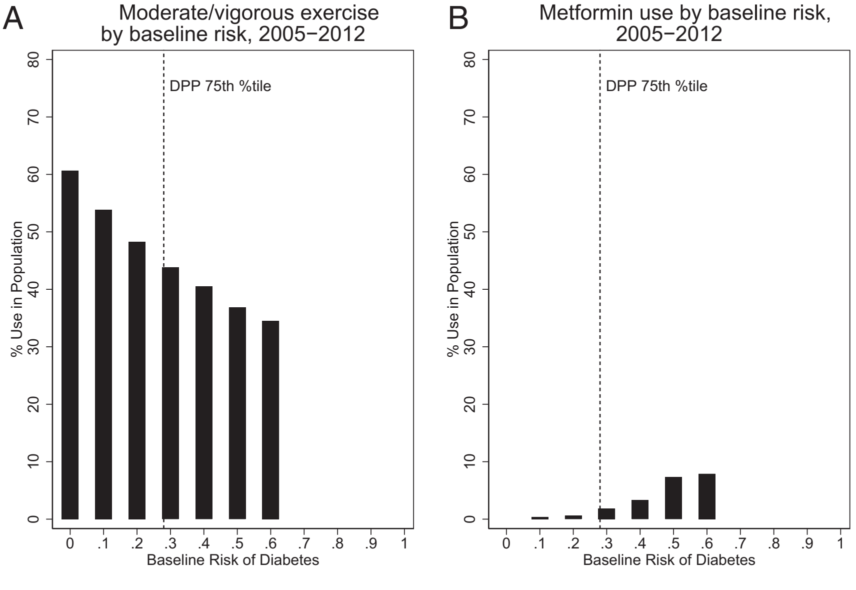
Studies have shown that metformin use and lifestyle modification are low among Americans with prediabetes (11). Implementation difficulties affecting low usage of the DPP's lifestyle treatment include the financial and organizational burden of designing such a large-scale lifestyle modification program. A significant barrier to large-scale adoption is the feasibility of translating intensive lifestyle intervention into real-world settings. The DPP lifestyle intervention cost US $2,780 per person over 3 y and required 135 visit hours (17). Although the monetary cost for the lifestyle intervention group was not significantly higher than that for the metformin group, the visit time was 3.5 times higher. The amount of visit time for lifestyle intervention is high relative to that for most services available in the current healthcare environment. This suggests that the shadow costs of using lifestyle intervention are higher compared to using metformin. More importantly, it alludes to the fact, although we do not have direct evidence to show, that the shadow costs for lifestyle intervention may be rising much faster over baseline risk than the shadow costs for metformin. It can be argued that if $\frac{\partial m_{\mathrm{OUT}}(\theta)}{\partial \theta}$ for lifestyle intervention $> \frac{\partial m_{\mathrm{OUT}}(\theta)}{\partial \theta}$ for metformin, then the use of lifestyle intervention may decrease over (θ) and we would not see a commensurate increase in the use of metformin in these patients, leading to higher than expected incidence of diabetes in patients with higher values of θ.

The implication of these results for disparities is straightforward. As long as two subgroups have different mass at higher values of θ at any point in time, we would expect the incidence of diabetes to be different in those two subgroups.

Empirical Uptake of Treatments Following DPP. We now examine the empirical uptake of treatments (adjusted for age and gender) following the DPP, given the theoretical insights regarding uptake discussed in the previous section. Fig. 5*A* shows the percentage of the at-risk population engaging in moderate or vigorous physical activity by race for the period 2005 to 2012, which is meant to proxy for the DPP trial's lifestyle modification treatment. The rates of physical activity by risk score appear to be decreasing at a swift rate, and this decline appears to be similar among all races (Fig. 6*A*). This decline is consistent with the expectations based on our theoretical model (Eq. **12**), which, recall, also suggests that this decline is conservative in comparison to what would have happened in an idealistic setting where a much steeper decline would be expected. Nonetheless, we observe a decline in the rate of use from about 60% at the lowest level of risk to about 35% at the highest risk where this rate could be reliably measured (i.e., sample size >30).

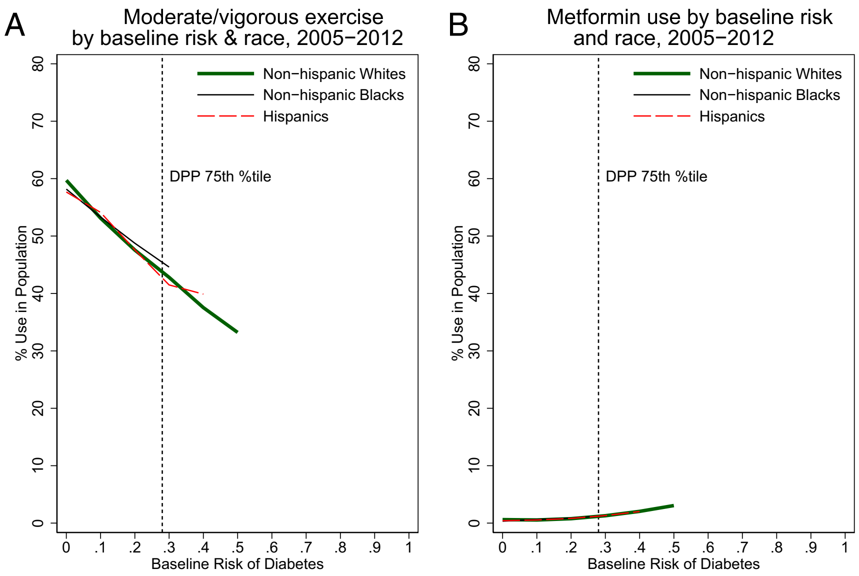
In comparison, the rate of metformin use in the time period 2005 to 2012, illustrated in Fig. 5*B*, appears to be low but rises over baseline risk, as expected in our theoretical model. Recall again that according to our theoretical model, this rise is expected to be conservative in clinical practice, compared to the ideal setting, where many more individuals would have used metformin as baseline risk increases. In fact, as conjectured above, the rise in metformin use over baseline risk comes nowhere close to bridging the decline in lifestyle intervention over baseline risk. Many social determinants of health plausibly influence treatment choices. The clinical baseline risk calculated in our case study depends on some of their proxies, such as history of high blood glucose, parental history of diabetes, height, and BMI. Naturally, we do see that the rates of treatment use rates vary over this baseline risk. Nevertheless, the fact that, after adjusting for age and gender, levels of treatment use do not vary across race (Fig. 6) within any baseline risk category indicates the role of other factors influencing treatment choices is considerably mitigated, at least in the current example.

The population rate of treatment use for each racial category can be obtained by integrating the risk-specific rates and outcomes over the respective distribution of θ. Since we have seen that over time non-Hispanic Blacks and Hispanics have larger densities of high θ values, one can predict that treatment choices, especially use of moderate/vigorous exercise, would decline more for non-Hispanic Blacks and Hispanics compared to Non-Hispanic Whites. But our model does not make any prediction about the implications for disparities in treatment choices. Indeed, when we look at the empirical data, we do see that use of moderate/vigorous exercise declines faster in non-Hispanic Blacks and Hispanics compared to non-Hispanic Whites from 2005 to 2007, with no concomitant differential



Adjusted for Age and Gender

Fig. 5. (*A* and *B*) Use of (*A*) moderate/vigorous physical activity and (*B*) metformin by risk score, 2005 to 2012.



Adjusted for Age and Gender

Fig. 6. (*A* and *B*) Use of (*A*) moderate/vigorous physical activity and (*B*) metformin by risk score and race, 2005 to 2012.

increases in the use of metformin (Fig. 7). Ironically, treatment disparities have declined as a result.

Implications for DPP Results for Disparities in Diabetes. Our model does make unambiguous predictions about the growth in disparities in realized outcomes across the racial categories. To summarize the role of the production and use of evidence in the context of the DPP and diabetes incidence, three main points arise: 1) The DPP trial likely enrolled a representative sample of the target population. However, over time, the distribution of baseline risk in the DPP trial no longer remained representative of the target population. This phenomenon particularly affected Hispanics and Blacks. 2) Average results from the DPP trial placed uniform emphasis on the superiority of lifestyle modification over metformin across all risk groups, even though metformin performed statistically similar to lifestyle modification in the higher quartile of the DPP risk distribution. 3) The shadow costs of lifestyle modification are higher than the shadow costs of taking metformin, and these differential shadow costs between the two treatments might be increasing over the baseline risk.

These observations demonstrate that the inadequate uptake of metformin for individuals at higher risk affected Hispanics and Blacks more due to the higher concentration of high risk in these groups compared to the non-Hispanic Whites. A back-of-the-envelope calculation reveals that had metformin increase in the higher-risk group been commensurate with the reduction of lifestyle intervention that is observed, then the potential reduction in disparity growth would have been[¶]

$$\begin{aligned}&\text{Reduction in disparity} = Pr(\text{At risk of prediabetes}) \times\\&\text{Incremental } Pr(\text{4th quartile risk}) \text{ for minorities} \times \text{Increase in}\\&\text{metformin use to compensate reduction in exercise at high}\\&\text{risk} \times \text{Absolute Risk Reduction for Metformin} \times\\&1{,}000 = 0.075 \times 0.085 \times 0.30 \times 0.22 \times 1{,}000 = 0.420.\end{aligned}$$

In other words, a targeted evidence generation and application in decision making could have decreased disparity growth by 10.5% (= 0.420/(7 − 3)). The most uncertain parameter in this calculation is the increase in metformin use to compensate reduction in vigorous exercise at high risk. There is uncertainty related to how the at-risk population would have responded to EBM if the EBM promoted risk-based treatment effects from the DPP. This remains to be an important area of future work.

Discussion

The effects of interventions and treatments are likely to be heterogeneous across patients because of how treatment exposures interact with patients' biologies and social environments. The current notion of evidence production for the effectiveness of treatments is stuck in the population-average context, i.e., what works best, on average, in a sample of patients. The implications of heterogeneous treatment effect most commonly discussed involve evaluating how representative the study sample is compared to the target population and how relevant the average result is to an individual patient choosing treatment. However, when these concerns are protracted over time, where a once representative study sample no longer remains representative of the target population as the underlying risk distribution changes, and as the individual's decision making is even more affected by relying on average results from a nonrepresentative sample, we fail to achieve the best health outcomes possible. To the extent that these phenomena affect different subgroups, such as race or socioeconomic status disproportionately, health disparities can grow over time.

The diabetes example highlights why we should think about moving toward some form of risk-based assessments of comparative effectiveness in our evidence generation infrastructure to facilitate decision making, not just at the completion of a research study but also to make sure that the results remain relevant over time. The diabetes example gives a conservative estimate on the effect on disparities, given the DPP was such a large trial and provided an adequate representation of the population distribution at the time of recruitment. For smaller nonrepresentative RCTs, growth in disparities may precipitate even earlier when following results of those trials.

We note here some limitations of our diabetes example. First, we use self-reports of moderate/vigorous exercise as a proxy for

[¶] Incremental Pr(fourth-quartile risk) for minorities: weighted average difference post-2005 from Fig. 4. Increase in metformin use to compensate reduction in vigorous exercise at high risk: from Fig. 5. Absolute risk reduction for metformin: from Fig. 3. The diabetes disparity, 7 − 3 = 4 (per 1,000 persons per year) is given by Fig. 2.

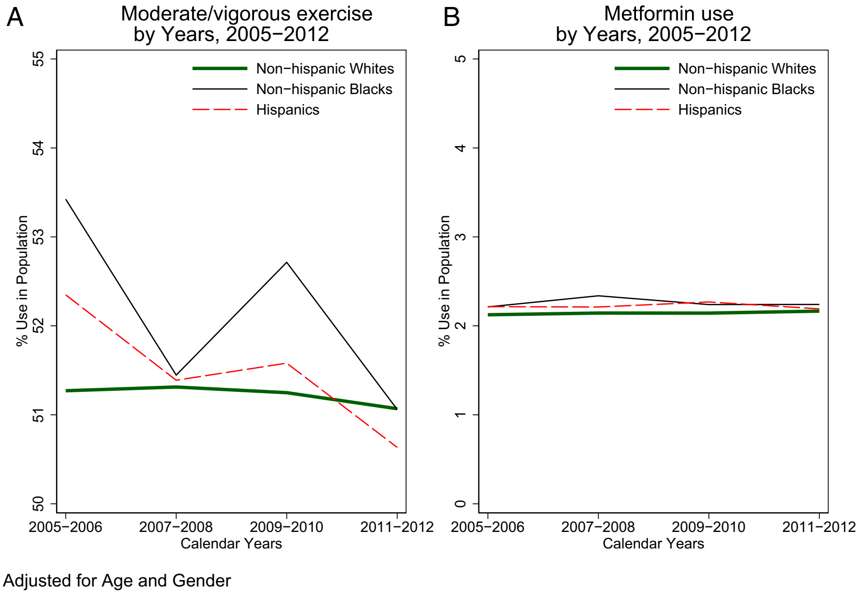


Fig. 7. (*A* and *B*) Use of (*A*) moderate/vigorous physical activity and (*B*) metformin by race over time.

intensive lifestyle modification intervention in the DPP. We do not know of any data that track the use of such an intervention nationally. However, the lifestyle intervention was designed as an intensive program with goals to achieve and maintain a weight reduction of at least 7% of initial body weight through a healthy low-calorie, low-fat diet and to engage in physical activity of moderate intensity, such as brisk walking, for at least 150 min per week. We use moderate/vigorous exercise as a proxy for the lifestyle intervention, recognizing that the shadow costs of engaging in such activities would be similar to or less than the shadow costs of engaging in the lifestyle intervention in its true form.

Second, we do not fully consider the dynamic effects of treatment selection in one period on the risk distribution of the at-risk population in the next period. In the case of diabetes, not choosing to use any of the risk-reducing treatments would have an ambiguous effect on the risk distribution of the remaining at-risk population in the next period. This is because, with no treatment, many currently high-risk individuals will develop diabetes and thereby move out of the at-risk pool for the next period. Similarly, with no treatment, many low- to moderate-risk individuals can move up to become high-risk individuals in the next period. However, this remains an area of active investigation, both in diabetes and in the theoretical development of such dynamics.

A key approach to resolving such a problem is to be able to develop prediction algorithms for individual-level treatment effect heterogeneity (18). Such algorithms can be constructed without identifying low-dimensional individualized characteristics such as genomic information, but rather by collapsing multi(high)-dimensional outcomes and behavior into individual-level latent characteristics, which can be used to establish individualized treatment effects. These prediction algorithms can be viewed as a hypothesis generation exercise at the individual level. However, these algorithms have two extremely useful implications for comparative effectiveness research. First, any attempt to individualize care based on prediction algorithms must begin with a hypothesis generation exercise, and therefore these results can provide valuable resources to clinicians and policymakers, who, in the absence of such resources, must rely on traditional comparative effectiveness results based on averages. The necessity of an algorithmic approach lies in the feasibility of translating enormous amounts of information to the bedside, without overwhelming physicians. Second, such individualized treatment effects will provide critical input to any confirmatory randomized trial evaluating and improving such prediction algorithms. Only by aligning the practical decision-making challenges to that of the evidence generation can it be ensured that we are obtaining the most out of scientific transnational research. Recently, a joint statement to promote the conduct of, and provide guidance for, predictive analyses of heterogeneity of treatment effects (HTE) in clinical trials has been proposed (19). We hope that future work will stress the importance of developing research designs, methods, and EBM frameworks, keeping these issues in mind.

Data Availability. Data used for the empirical analysis in diabetes are based on original and derived variables from the NHANES datasets. Original data can be found at https://www.cdc.gov/nchs/nhanes/index.htm. The analytical Stata dataset and a variable dictionary are available at the Open Science Framework: https://osf.io/gr29s/files/.

ACKNOWLEDGMENTS. We are grateful for comments from Charles Manski, Dave Veenstra, two anonymous referees and the editor, and seminar participants at the University of Washington and the University of Southern California. All opinions and errors are ours.

1. C. F. Manski, Minimax-regret treatment choice with missing outcome data. *J. Econom.* **139**, 105–115 (2007).
2. D. Kravitz, N. Duan, J. Braslow, Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.* **82**, 661–687 (2004).
3. C. Worsham, A. B. Jena, The art of evidence-based medicine. *Harvard Business Review*, 30 January 2019. https://hbr.org/2019/01/the-art-of-evidence-based-medicine. Accessed 9 March 2020.
4. T. Philipson, The evaluation of new health care technology: The labor economics of statistics. *J. Econom.* **76**, 375–395 (1997).
5. A. Malani, Identifying placebo effects with data from clinical trials. *J. Polit. Econ.* **114**, 236–256 (2006).
6. Diabetes Prevention Program Research Group, Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med.* **346**, 393–403 (2002).
7. L. Geiss *et al.*, Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. *J. Am. Med. Assoc.* **312**, 1218–1226 (2014).
8. E. Ely *et al.*, A national effort to prevent type 2 diabetes: Participant-level evaluation of CDC's national diabetes prevention program. *Diabetes Care* **40**, 1331–1341 (2017).

9. A. D. Roy, Some thoughts on the distribution of earnings. *Oxf. Econ. Pap.* **3**, 135–146 (1951).
10. A. Basu, Welfare implications of learning through solicitation versus diversification in health care. *J. Health Econ.* **42**, 165–173 (2015).
11. J. B. Sussman, D. M. Kent, J. P. Nelson, R. A. Hayward, Improving diabetes prevention with benefit based tailored treatment: Risk based reanalysis of diabetes prevention program. *BMJ* **350**, h454 (2015).
12. Diabetes Prevention Program Research Group *et al*., The diabetes prevention program: Baseline characteristics of the randomized cohort. *Diabetes Care* **23**, 1619–1629 (2000).
13. Agency for Healthcare Research and Quality, Rockville, Screening and interventions for overweight and obesity in adults. https://www.ahrq.gov/ncepcr/tools/healthier-pregnancy/fact-sheets/obesity.html. Accessed 1 June 2020.
14. American Diabetes Association, The prevention or delay of type 2 diabetes. *Diabetes Care* **26** (suppl. 1), s62–s69 (2003).
15. D. M. Kent *et al*., Risk and treatment effect heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. *Int. J. Epidemiol.* **45**, 2075–2088 (2016).
16. American Diabetes Association *et al*., The diabetes prevention program. Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care* **22**, 623–634 (1999).
17. DPP Research Group, Costs associated with the primary prevention of type 2 diabetes mellitus in the diabetes prevention program. *Diabetes Care* **26**, 36–47 (2003).
18. A. Basu, Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *J. Health Econ.* **30**, 549–559 (2011).
19. D. Kent *et al*., The predictive approaches to treatment effect heterogeneity (path) statement: Explanation and elaborations. *Ann. Intern. Med.* **172**, 35–46 (2020).