Data-heavy "wellness rewards" programs allow covert discrimination based on race, social class, and pre-existing conditions

Kaiser Health News recently put out an article about wellness rewards programs that ask for information on disease history, seat belt use, and groceries. This is pretty invasive, and the article focuses on privacy risks, which are substantial. Aside from privacy, though, there is another absolutely crucial issue that they don't mention:

It is impossible to implement grocery-based wellness rewards programs without implicit discrimination, even assuming companies act in good faith.

That's what this post is about. I will touch on the potential for bias based on **social class** and **pre-existing conditions**, but I mostly focus on **race**. Before we get into the details, remember that discrimination has arisen time and again through supposedly neutral decision criteria. These topics are controversial, and you may not agree with every single claim people make, but examples abound, ranging from redlining, predictive policing, and drug policy to IQ tests and college admissions. This is a common pattern.

Furthermore, grocery purchases offer incredibly detailed information. You can infer race, class, and even pre-existing conditions from someone's groceries -- not perfectly, but much better than chance. (I assume groceries could be monitored with a low rate of people scamming the system; otherwise the topic is moot.) Because of this, even so-called "fairness through unawareness" -- setting rebates based on groceries and health outcomes with no knowledge of race, class, or pre-existing conditions -- will still result in unfair penalties.

Grocery purchases correlate with race and class

It is common sense that grocery purchases reflect social class. The specifics are almost painful to list: is your customer buying filet mignon and caviar, or beans, rice, and Kraft dinners? Race is not quite as obvious, but according to market research by Nielsen, simple rules of thumb based on purchases of branded and unbranded produce, meat, and seafood could reveal information about race.

More relevant to my point here is that healthy eating scores produced by well-meaning health researchers, with no intent to discriminate, still find strong racial disparities. For an example regarding race, I looked at this study.

> Li, W., Youssef, G., Procter-Gray, E., Olendzki, B., Cornish, T., Hayes, R., ... & Magee, M. F. (2017). Racial differences in eating patterns and food purchasing behaviors among urban older women. *The*

> *journal of nutrition, health & aging*, *21*(10), 1190-1199.

Some back-of-the envelope calculations on this small Washington D.C. study indicate that groceries could separate Black and White participants with about 68% accuracy (versus 50% by chance), subject to certain assumptions (see appendix). For those wanting evidence on a broader scale, with more geographic and topical diversity, a review of more than 100 nutrition studies explains:

> Accessing healthy food is a challenge for many Americans—particularly those living in low-income neighborhoods, communities of color, and rural areas.

Any honest attempt at grocery-based wellness rewards has a moral and legal responsibility to address these issues. Below, I make the case that this is impossible in theory given the state of the art about how to define "fairness" for a predictive algorithm, and impossible in practice given our incomplete understanding of health science.

On fairness through unawareness

Some people would object to this argument, saying that healthy eating scores are fair as long as they properly reflect biological impact of diet on health. Why bring in complicated considerations about race, class, and pre-existing conditions when we could be having a simple discussion about rewarding people who make healthy choices of what to eat? There are several reasons.

First, causal effects in nutrition research are *notoriously* difficult to parse out. Diet is complex and difficult to measure, and health effects of diet overlap confusingly with geographic and social trends. Even obvious patterns are compatible with multiple different biological mechanisms, and these would often suggest different healthy eating criteria. For instance, colon cancer rates are higher in developed countries, and they have risen rapidly with recent industrialization (source). Is this due to under-reporting in less wealthy places? Lack of screening? Genetics? Proportion of young versus old people? Exercise habits? Consumption of fiber? Consumption of animal fat? Carcinogens present in cooked meat? Viruses more often present in raw beef? All of these explanations are plausible, but asking people to eat more fiber is different from asking them to eat less animal fat or to eat less beef or to exercise. The details will matter to a carefully done wellness rewards program.

Second, try to enter a room with a Black consumer and a White consumer and explain why it's fair that the foods one of them grew up eating will now cost her extra in health insurance. If they object, their intuition matches a carefully reasoned criterion called counterfactual fairness. To be fair according to this work, predictions must remain unaffected by changes in someone's racial background, including their own race, their family's races (see appendix 2), and anything that depends on those attributes (see appendix 3). Assessing whether a family's food preferences depend on race would require a thorough family history project full of impossible counterfactual questions. Would your great-grandmother have borrowed that cookbook from her neighbor if she were White? Would she have even lived in that same neighborhood? Gone to that same

market? Had the same amount of money to spend on food? Given the uncertainties involved, the only feasible fair practice is to avoid using groceries for wellness rewards.

Third, health research -- even completely reasonable and well-meaning health research -- is affected by our perceptions about race. For example, consider the uncontroversial idea that some populations are less healthy in some ways. This idea affects the following example, which is a large, geographically widespread, multi-racial study of the associations between race, geography, Southern-style diet, and stroke risk.

> Judd, S. E., Gutiérrez, O. M., Newby, P. K., Howard, G., Howard, V. J., Locher, J. L., ... & Shikany, J. M. (2013). Dietary patterns are associated with incident stroke and contribute to excess risk of stroke in black Americans. Stroke, 44(12), 3305-3311.

This is a careful and impressive piece of work. It does a great job at its stated purpose. But, it prioritizes diet-associated disease in Black Americans, leaving other important questions unanswered. The first sentence of the paper: "Black Americans and residents of the Southeastern United States are at increased risk of stroke."Ignoring other heatlh outcomes can lead to strange findings. Another unhealthy eating pattern that the authors name "Sweets/Fats" was more common in White Americans. Despite summarizing items such as candy and desserts, "Sweets/Fats" consumption correlated with a *reduction* in stroke risk. The authors did not expect this, and their best explanation is that eating Sweets/Fats "protects" people from strokes only by killing them with cancer or heart disease before a stroke can happen.

Any wellness rewards program that accounts for this study could take its well-meaning focus on Black Americans' stroke risk and use it to impose higher health insurance prices, while essentially ignoring unhealthy choices common among White Americans. Again, the problem is not the study. We have to accept that, through no fault of the authors, our understanding of nutrition and health is incomplete, and it's incomplete in ways that are deeply affected by race. Under these circumstances, designing a fair wellness rewards program is not feasible.

Grocery purchases also correlate with pre-existing conditions

Many readers will not be surprised that, in aggregate, dietary preferences and healthy eating indices differ by race and class. More surprising to me is that groceries can be used to detect pre-existing conditions on an individual level. For example, Brown University economist Emily Oster recently used grocery purchases to infer diagnosis of diabetes for a study unrelated to wellness rewards. She started with only directly relevant items (such as glucose testing products), but found that adding information from regular groceries substantially improved the false discovery rate (from 68% to 40%). This finding shows that in principle, routine grocery purchases may contain substantial individual-level information on pre-existing conditions. Other pre-existing conditions include AIDS, pregnancy, obesity, alcoholism, and kidney failure. How well could these be predicted based on groceries?

Companies are already finding indirect ways to screen for pre-existing conditions, for example by putting AIDS

drugs on tiers with very high copays (source). It might be difficult to take wellness rewards as a similar opportunity, though. If the wellness reward is structured as a rebate, like the name suggests, then they would have to advertise the AIDS-patient price to everyone and then selectively give out large rebates. Even customers whose groceries qualify for the maximum rebate would probably not tolerate this practice. Furthermore, if the process were transparent, it could generate a big backlash. The most realistic scenario is probably either no discrimination against pre-existing conditions, or subtle discrimination that imposes only a fraction of the extra price and targets people in a messy way, hitting many healthy people and missing many who do have pre-existing conditions.

Summary

If we are going to charge some people more for health insurance, we have a moral and legal responsibility to ensure we are not penalizing them based on race, class, and pre-existing conditions. This is impossible: grocery purchases are wound up in our lives much too tightly for it to work. Market research indicates that simple rules of thumb based on purchases of branded and unbranded produce, meat, and seafood could reveal information about race, and health studies find that even carefully designed healthy eating scores stratify people by race. Social class is similarly vulnerable. Emily Oster showed that grocery bills could be used -- not perfectly, but much better than random -- to ferret out people with a pre-existing condition (diabetes). For complex prediction systems based on grocery purchases, avoiding discrimination is impossible in practice.

Even if well-meaning programs focus on causal effects of diet, "fairness through unawareness" is not enough to avoid unacceptable bias. On principle, people should not be penalized for cultural preferences, even those with a genuine causal effect on health. On the practical side, diet is too difficult to disentangle from other determinants of health, and the legacy of racism is baked into well-meaning health research, shaping whose diet and what health outcomes are scrutinized.

Call to action

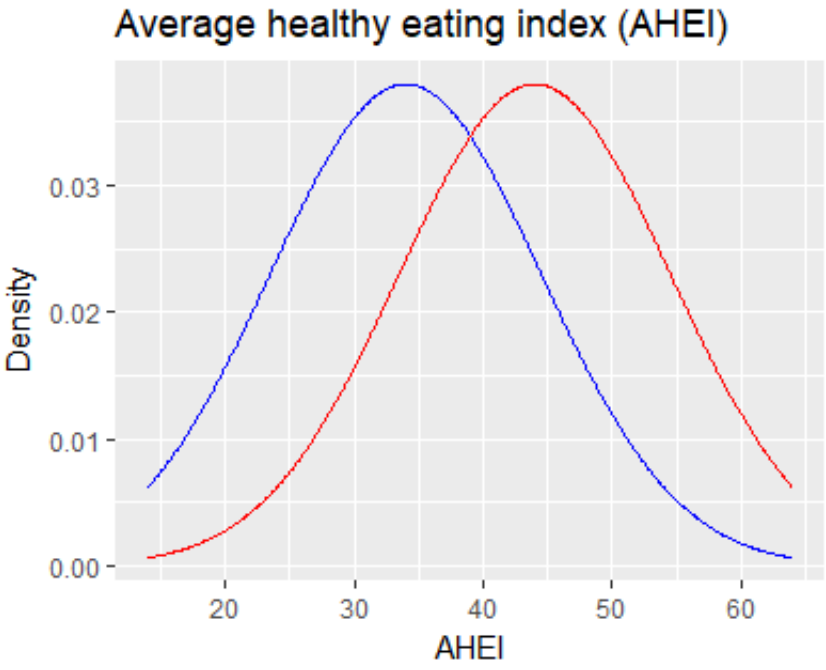
At the very least, wellness rewards programs should make their datasets, methods, and principles transparent. They should engage with these issues and make the case to consumers for why their practices are fair. But ideally, health insurers should abandon grocery-based wellness rewards. Perhaps they could instead focus their efforts on making healthy food accessible and convenient for their members, or lobby for extending agricultural subsidies to include healthier options.

In the Houston incident relayed by Kaiser Health News, the city government switched to a separate program after a backlash. The KHN article cites objections from the police union and other city employees as key to that decision. Labor unions are strong enough to negotiate with large employers about institutional decisions on health benefits, and the best way for consumers to regain power over their data is to get organized.

Appendix

Appendix 1: Accuracy of Li et al's healthy eating index as a Black-White classifier

Li et al. includes an overall Average Healthy Eating Index, for which the higher-scoring group ($p < 0.001$) was White participants. Li & coauthors report group means of 44 and 34 points with standard deviations of 10.5. (I'm rounding the numbers.) If the scores were perfectly Normally distributed, it would give an overlap that looks like this.



In Washington D.C., where Li & coauthors' study recruited, Black and White populations each make up roughly 45% of the city, so it's fair to put the same area under the curves even though that wouldn't make sense nation-wide. Dropping a decision boundary midway at 39 gives a Black-White classifier with about 68% accuracy, compared to 50% by chance.

The assumption of normally distributed scores is not reasonable, and it could bias the estimate in either direction, but when I replace the normal distribution with a right-skewed Gamma distribution having the same mean and sd, the numbers hardly change. A bigger issue: it's unclear how statistical trends will change when expanded from ~100 older women in DC to national-level or regional-level wellness vendors. This is why I cited the Food Trust review article as well.

Appendix 2: On counterfactual fairness

Here is the counterfactual fairness paper by James Kusner and colleages.

> Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4066-4076).

They advocate for the following definition of "Counterfactual Fairness". Y represents the variable being predicted -- for instance, yearly healthcare costs. X represents the data used to make the prediction -- for instance, grocery purchases. A represents a protected attribute such as race or gender. $\hat{Y}$ is the prediction of Y. $\hat{Y}_{A \leftarrow a'}$ represents a counterfactual prediction for someone whose race and/or gender has been magically altered. For all possible values of x, a, and a', the criterion demands

$$Pr(\hat{Y}|X = x, A = a) = Pr(\hat{Y}_{A \leftarrow a'}|X = x, A = a)$$

.

In plain English, if you alter someone's gender identity or racial/ethnic background (for example), it should not affect the predictions made about them. Following causal statistics, I'll call this an "intervention." For biological/physical sex, the intervention might be to roll back time to when someone is conceived and replace the paternal X chromosome with the paternal Y chromosome or vice versa.

A key aspect of this definition is that you must allow the intervention's effects to propagate. For the case of biological/physical sex, the intervention happens at conception, and a fair prediction should be invariant to all the resulting changes in either life experience or innate characteristics.

For a concept like race, which is not biological but rather tied up with culture, history, and perceptions of others, it is not clear to me what the intervention is. After studying the article, I conclude that it would be far-reaching. Kusner et al specify that the set of protected characteristics needs to be ancestrally closed with respect to the underlying causal graph. In plain English, that means if race is protected, so are "mother's race", "father's race", and so on. This has massive implications when applying the criterion to race as a protected attribute. If you want your prediction system to meet Kusner et al's standard of racial equity, you should imagine the intervention happening generations ago, with all the cascading experiences of racial identity over centuries. Your predictions should still come out the same.

Appendix 3: More on counterfactual fairness

Why do we have to completely avoid using diet in our predictions, just because it has been affected by racial / cutural history? In the counterfactual fairness paper, Lemma 1 and the discussion surrounding it answers this question. The only practical way to avoid flunking the criterion is to avoid variables that are downstream of protected attributes in the causal graph. Although this is not mathematically necessary in situations where the causal model is completely specified, it's necessary in real life, where the causal model is murky.