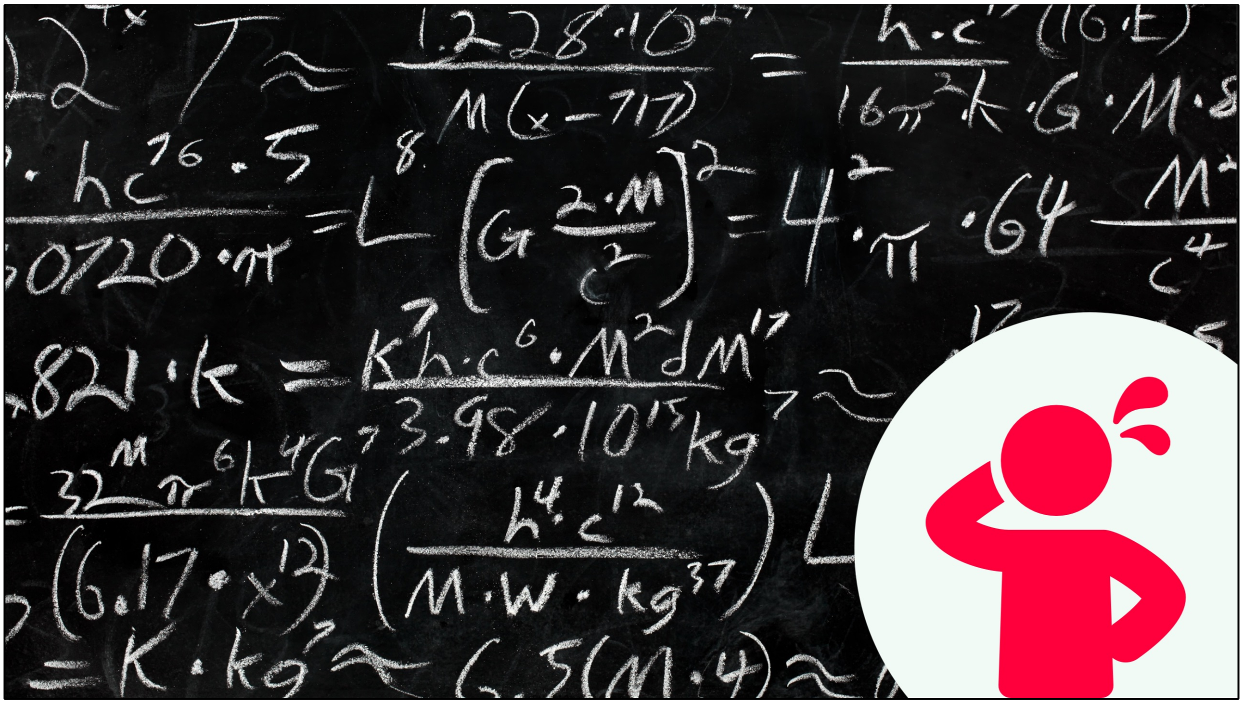
Missingness in Medicine

Addressing the messiness of healthcare data

Carl Preiksaitis, MD
Department of Emergency Medicine
Stanford School of Medicine

Stanford MEDICINE | Emergency Medicine

Hi, my name is Dr. Carl Preiksaitis, I'm a clinical instructor of emergency medicine and medical education fellow at Stanford. Today I'm going to do a brief overview of the issues surrounding "missingness" in healthcare data.



Disclaimer: I am not a statistician. I can't go into detail about methodologic approaches to address problems with your datasets. My goal is to introduce you to this problem of missing data and give you some things to consider when working with data collected from patients.

MIA
Missingness in Action

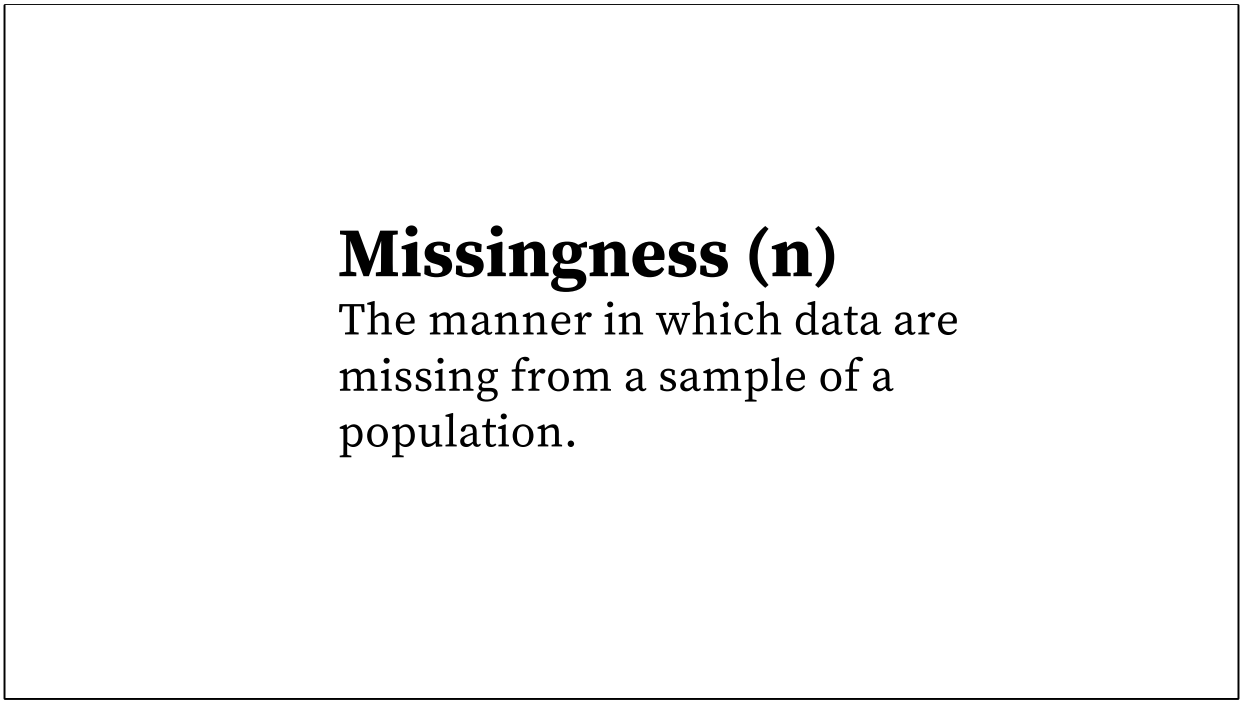
September 22, 2022
A Stanford Conference on the
Absence of Data and
the Future of AI in Health Care

StanfordMIA.org

Stanford MEDICINE | Emergency Medicine

HAI Stanford University Human-Centered Artificial Intelligence

The information in this talk is largely informed by a recent conference that we put on at Stanford called “missingness in action.” So, sincere thank you to Dr. Christian Rose and the MIA team and speakers for a lot of the information I’m going to touch on today. If this topic is of interest to you, you can review the recordings of the conference at stanfordmia.org



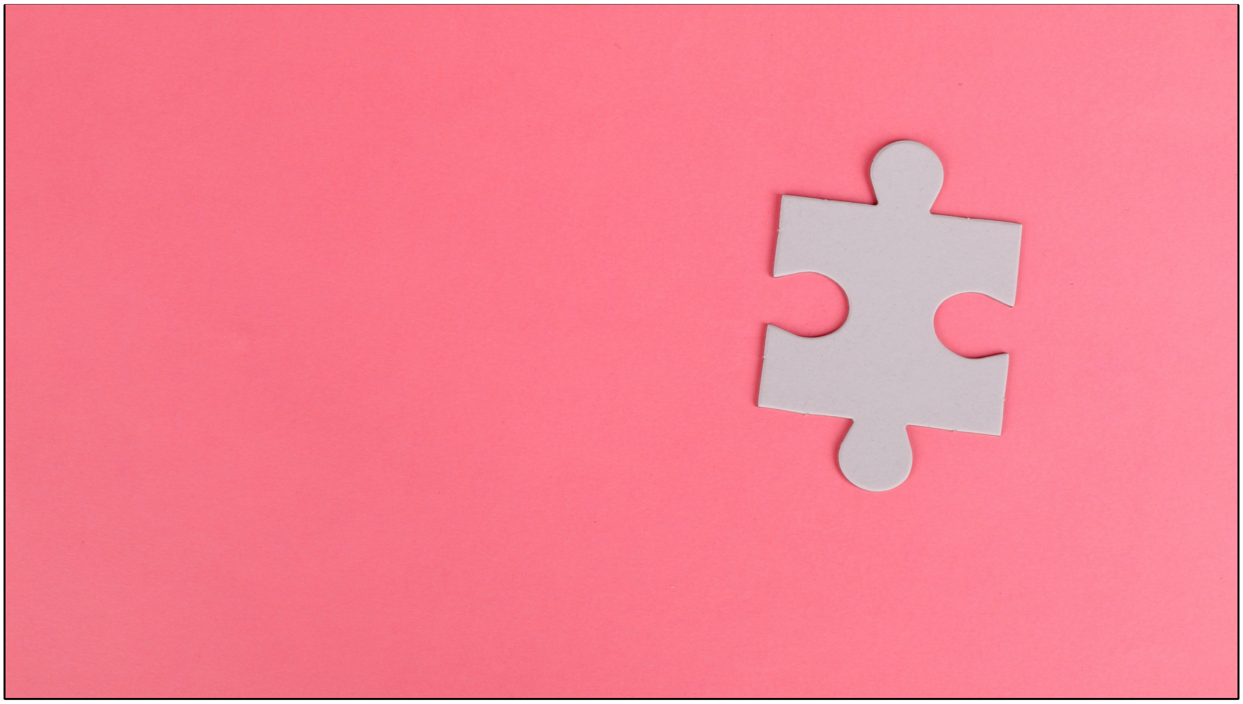
Missingness (n)

The manner in which data are missing from a sample of a population.

First let me start by defining missingess in data. One definition is the manner in which data are missing from a sample of a population. Missingness in healthcare data then is a description of how and in what ways a given data set fails to represent the patient population under study.

When we consider advances in how we use healthcare data, specifically applications like artificial intelligence or machine learning, the success of these interventions are dependent on the data they use for analysis.

But as healthcare data grows in volume and verocity, the more we measure, the more we realize that more data doesn’t necessarily mean better data



We encounter gaps in our data, as well as over-and under-representation of conditions or entire populations presenting several problems. How do we control for missing information, how do we identify and fill the gaps this missing data presents, and how do we plan for incorporation of data we have not yet measured?

Let me first get deeper into some definitions of missing data.

SMD

STRUCTURALLY MISSING DATA

There are several statistical categorizations for missing data. The first is structurally missing data. This is data that is missing for an apparent reason. An example of this would be looking at pre hospital seizure management and one ambulance company doesn’t stock midazolam. So an entire group of patients would be missing information in the time to midazolam. This is easy to predict and address for in analysis.

MCAR

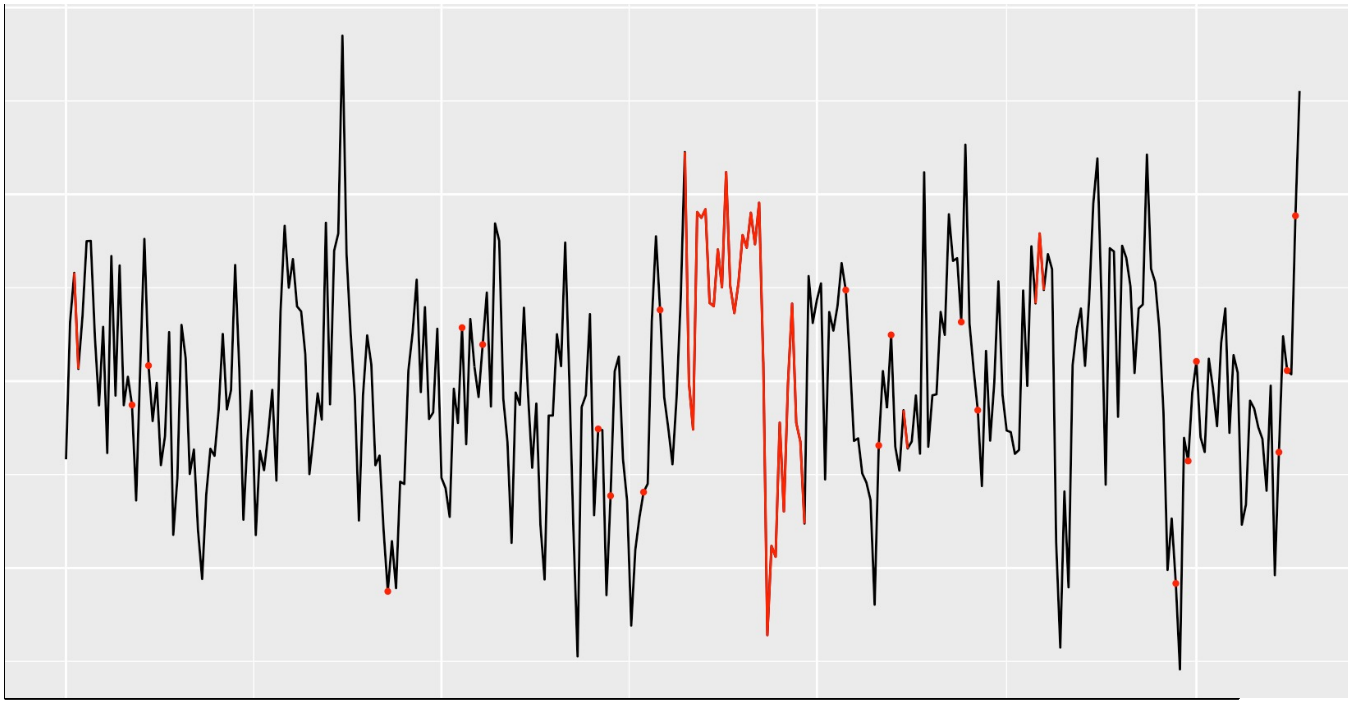
MISSING COMPLETELY AT RANDOM

Missing Completely at Random (MCAR) is the next category. This occurs when the probability of data being missing is unrelated to all observed and unobserved characteristics. The example would be administering a survey and some are lost in the mail. Since the missingness is completely random, the data can be analyzed without any adjustment, however this situation is rarely plausible with clinical data.

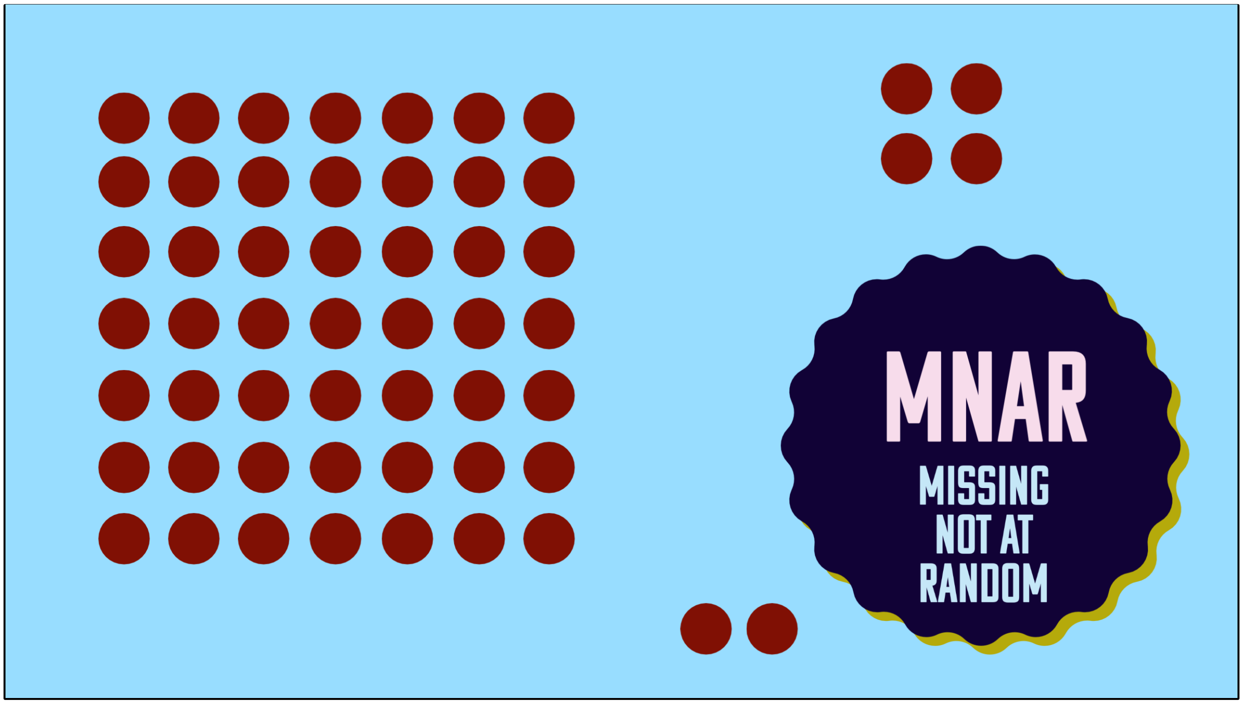
MAR

MISSING AT RANDOM

The third category is Missing at Random (MAR). This occurs when missing values are fully dependent on observed values. The example would be that 90 day outcomes are missing from the sickest patients in your study because they died.



There are ways to impute or infer values. One common method is multiple imputation, but there is no method available to prove that data is missing at random. It requires assumptions on the researchers to make these adjustments.



MNAR

MISSING NOT AT RANDOM

Finally, the most difficult to deal with is data that is Missing Not at Random (MNAR). This occurs when missing values are dependent on unobserved or unknown factors. Ie: the reason a data point is missing is directly related to the data point under scrutiny. Statistical adjustment is not possible in this case and it’s difficult to determine when data is missing not at random.

Unfortunately healthcare data is often missing “not at random,” which severely limits the predictive ability of tools like machine learning and AI. Particularly this limits the clinical applicability for those patients who may need them the most.

mHealth

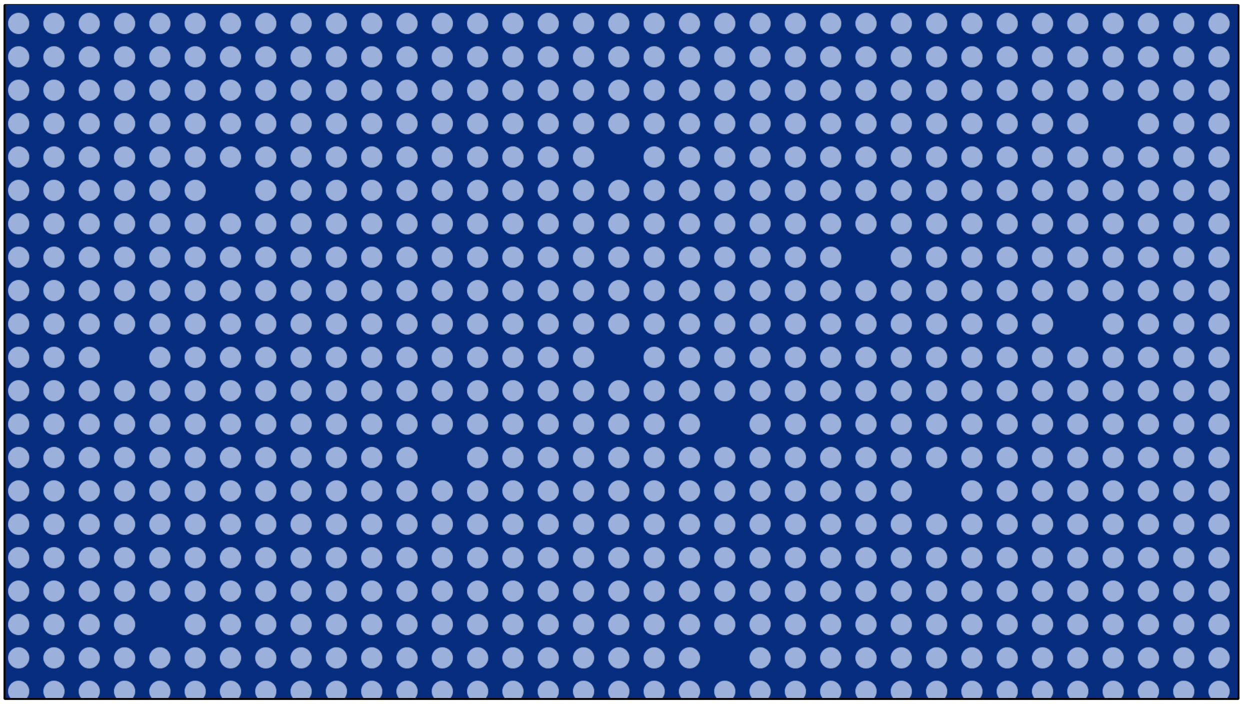
As a particular case example, MNAR (missing not at random data) is particularly problematic in mobile health interventions. Specifically, mobile health has a very high attrition rate, and hence a lot of missing data. However, this data is generally linked to the effectiveness of an intervention. That is, people are more likely to drop out if the intervention wasn’t effective for them.

Now, there are not adjustments per say to determine whether this affects your results, but the current best recommendations are to perform some sensitivity analysis to determine how big an effect the missing data has on your results. Again we are forced to make assumptions. Do we assume that the intervention failed for everyone who dropped out? This is certainly the most conservative (and is often what is done in addiction research), but adopting such a strong assumption can dilute the real benefit of an intervention.

Unfortunately the problem grows in magnitude when considering working with larger volumes of healthcare data, which may have gaps reflective of systems level issues.

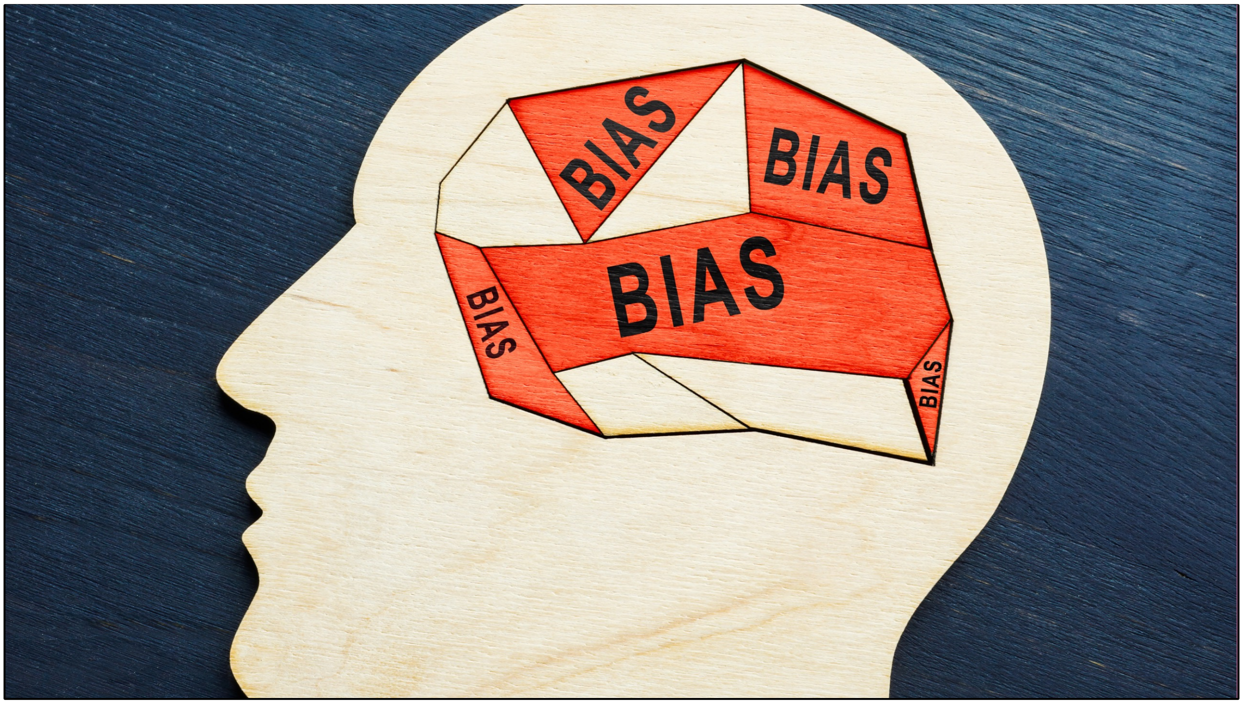
Currently, even our largest clinical data sets contain information on only a small fraction of the population and represent those communities who have historically benefitted from the greatest access to healthcare. The resulting imbalances and lack of representation of systematically oppressed populations may be the most impactful limitation of medical data and its generalizability. These populations rarely make it into our observations and calculations not because of a lack of need, but rather because the medical community has rarely examined these data in a way that acknowledged the various needs and expectations of these groups. Thus, our observations often represent only a small fraction of the truth.

I think this is the most important observation to make. The volume of data we have available to us gives us the false impression that this data is totally representative of a given population. We must be specific and deliberate in identifying the population our datasets apply to.



This will allow us to appropriately limit the application of our results and identify areas where we need more or higher quality data.

It can be challenging, however, to identify these areas of missing information when looking at a large data set. One way to address this is to carefully roll out any applications with close scrutiny of the individual, patient-level effects of these interventions. To do this successfully, we need better infrastructure to appropriately study populations that often fall outside of the current healthcare system.

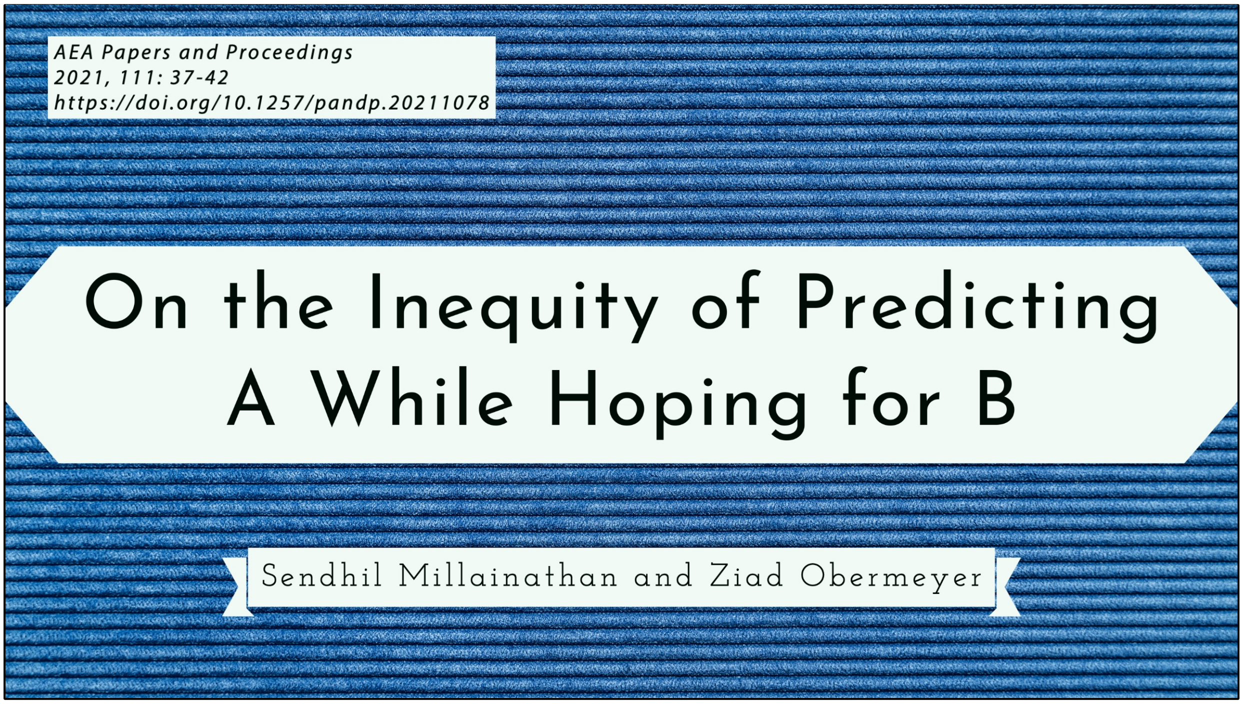


If you talk to any statistician about missing data, the best solution is to avoid it and collect more data.

But even if we recruit data from a greater subset of patients, these data are a product of a number of different, potentially biased factors that are often difficult to detect and account for. Looking at race is a powerful example of this. Even if algorithms have no inputs related to race, they still can be racially biased.

LABEL CHOICE BIAS

One way that this can occur is label choice bias. It arises when we select a biased proxy for an algorithm's prediction target. There are lots of examples of this. We incentivize teachers to improve test scores and we get higher scores, but not necessarily more learning. We pay hospitals to deliver treatments and we get more utilization, not healthier patients.

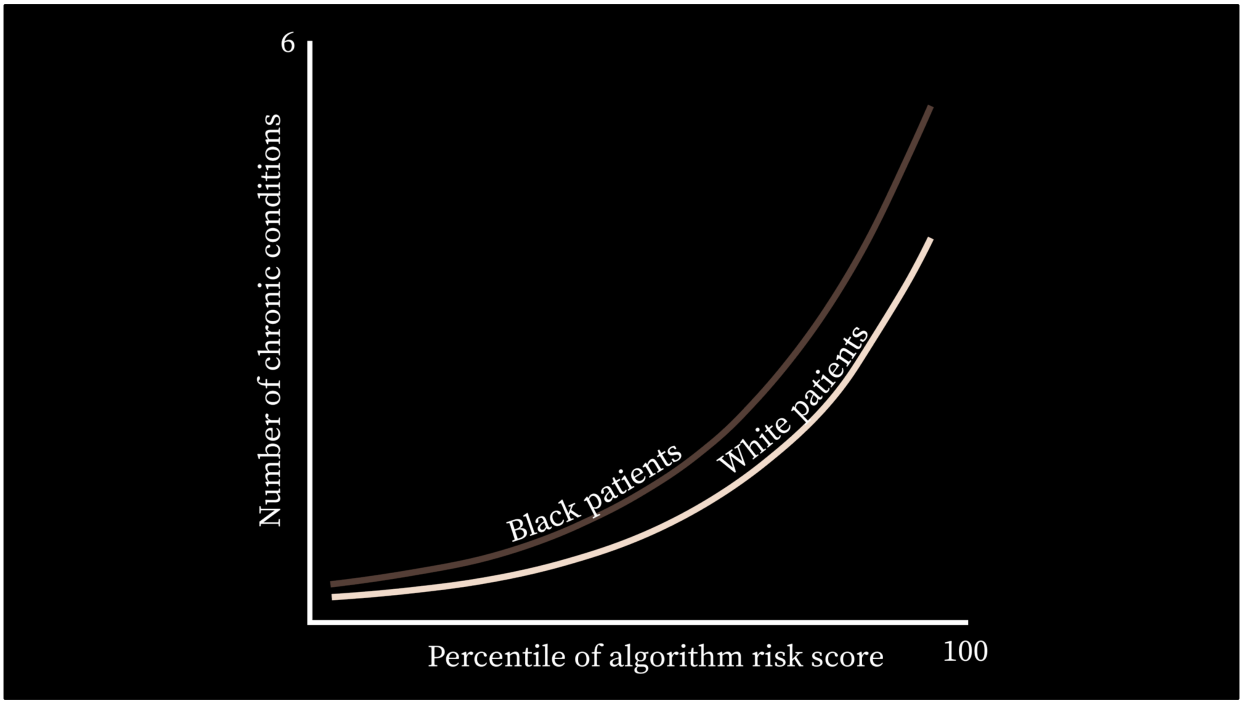


AEA Papers and Proceedings
2021, 111: 37-42
https://doi.org/10.1257/pandp.20211078

On the Inequity of Predicting A While Hoping for B

Sendhil Millainathan and Ziad Obermeyer

The example I want to reference is from an excellent paper from Mullainathan and Obermeyer published last year in Papers and Proceedings of the American Economic Association. They reference an algorithm developed to target extra help to patients with complex medical needs. Most healthcare systems use “high-risk management” programs to help high-risk patients manage chronic illnesses. Because resources of these programs are costly, algorithms are often used to help allocate funds to those who need them most.



This algorithm explicitly excluded race, but when the researchers stratified the results of the algorithm based on reported race of the participants, the algorithm repeatedly scored white patients at higher risk for a similar number of chronic illnesses.

Where did the algorithm (that purposefully excluded race from its calculations) go wrong? It becomes clearer through a close examination of the labels chosen to predict healthcare utilization. The developers of the algorithm chose healthcare spending as a proxy for healthcare needs, however on reflection we can see that there is a large issue with this assumption. Healthcare spending is not agnostic to race or socioeconomic status, and we historically know that for similar rates of chronic disease, black patients will spend less on healthcare due to issues with access and economic means.

This is just one powerful example of how algorithms can fail us if we do not recognize the larger context the data we analyze are situated in. To that end, any data-driven solution to a healthcare problem must be developed in concert with a nuanced understanding of all factors of the system which may or may not be represented in the data.

We must also re-examine the effects and inputs of our algorithms to look for evidence of missing data or label choice bias that can have unintended consequences.

In summary, we need to be cognizant of the missingness and messiness of healthcare data. There are lots of ways data can be missing and some ways to address this statistically, but when data is missing not at random or affected by faulty label choice, it can be difficult to predict how any intervention built on that data will perform.

When we develop new solutions based on existing data we must realize that these may not apply to all patients or populations. Because of this we must scrutinize the patient-level effects of any intervention and monitor for unintended consequences. To do this successfully, we need better healthcare infrastructure for marginalized and excluded populations. Not only will this aid in higher quality data collection, but it will ensure that what is built upon the data is done so ethically and responsibly.

Stanford MEDICINE
Emergency Medicine
EMERGENCY
MARC & LAURA ANDREESSEN EMERGENCY DEPARTME

@CMPreik

f