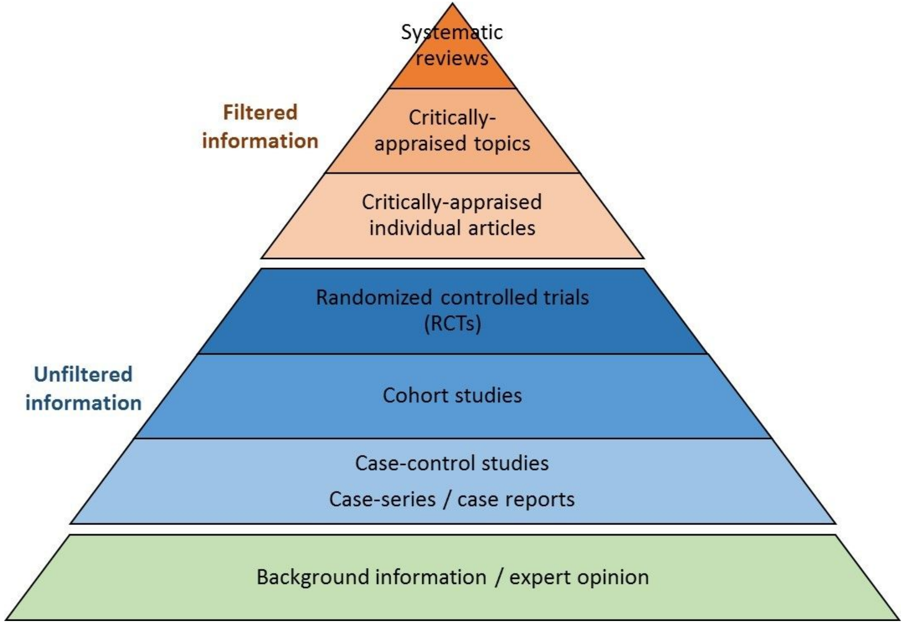
Chapter 14: Bias and the Electronic Health Record

Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 30 min, max) while following along in the notes



Systematic reviews

Critically-appraised topics

Critically-appraised individual articles

Randomized controlled trials (RCTs)

Cohort studies

Case-control studies

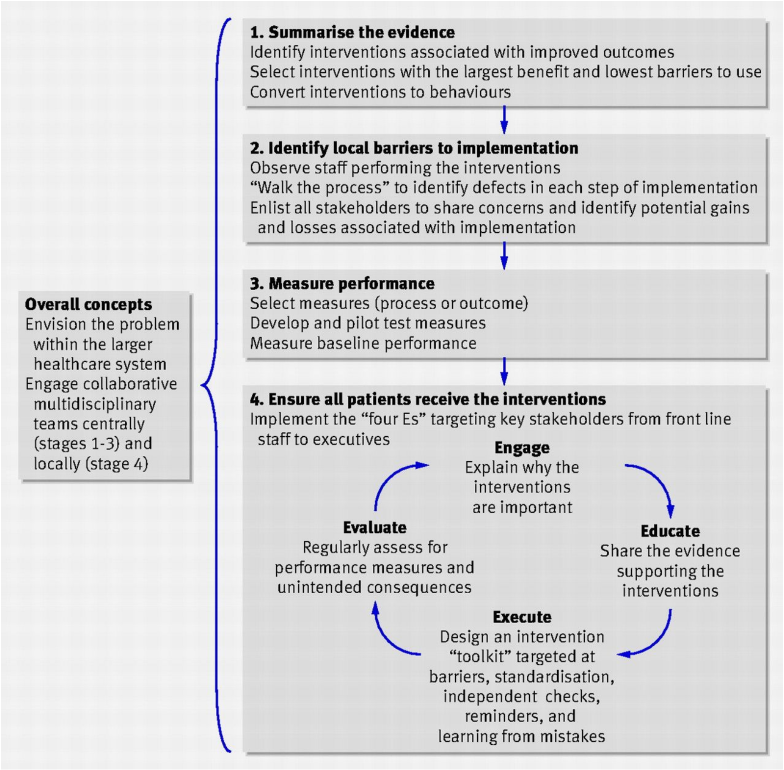
Case-series / case reports

Background information / expert opinion

Filtered information

Unfiltered information

https://latrobe.libguides.com/ebp/study-design



Overall concepts
Envision the problem within the larger healthcare system
Engage collaborative multidisciplinary teams centrally (stages 1-3) and locally (stage 4)

1. Summarise the evidence
Identify interventions associated with improved outcomes
Select interventions with the largest benefit and lowest barriers to use
Convert interventions to behaviours

2. Identify local barriers to implementation
Observe staff performing the interventions
"Walk the process" to identify defects in each step of implementation
Enlist all stakeholders to share concerns and identify potential gains and losses associated with implementation

3. Measure performance
Select measures (process or outcome)
Develop and pilot test measures
Measure baseline performance

4. Ensure all patients receive the interventions
Implement the "four Es" targeting key stakeholders from front line staff to executives

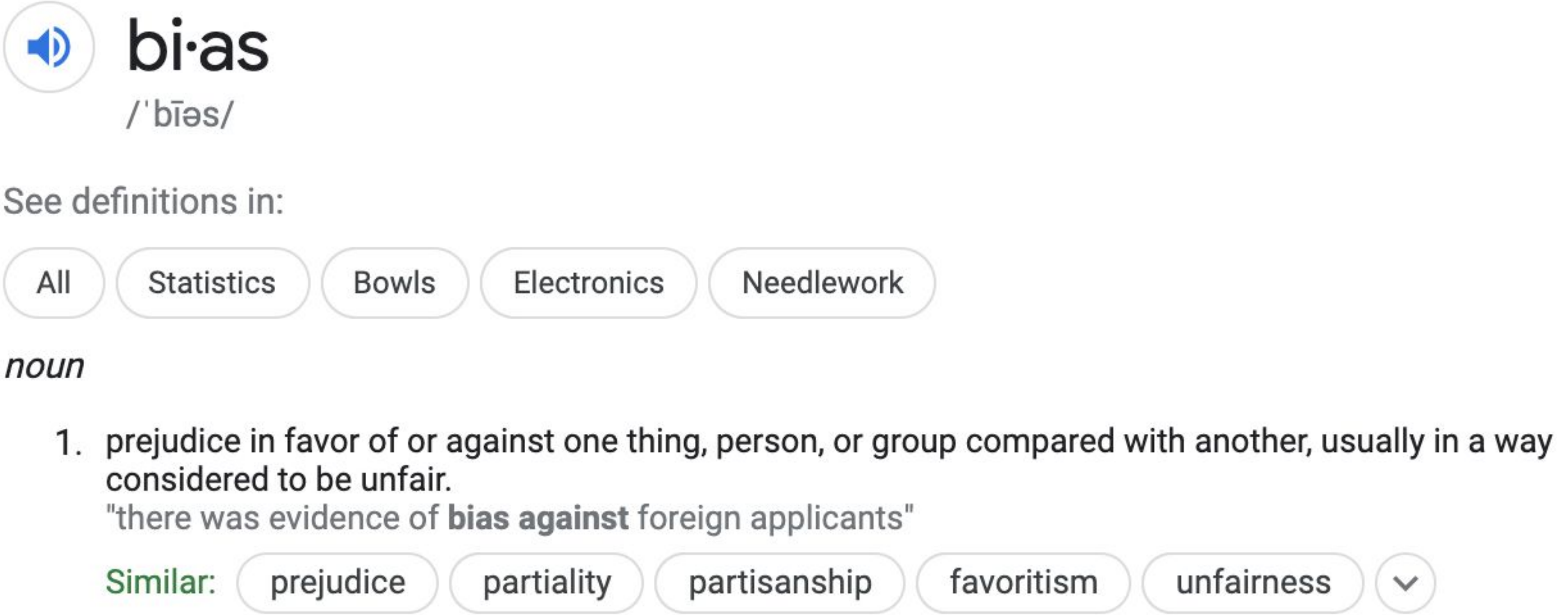
Engage
Explain why the interventions are important

Educate
Share the evidence supporting the interventions

Execute
Design an intervention "toolkit" targeted at barriers, standardisation, independent checks, reminders, and learning from mistakes

Evaluate
Regularly assess for performance measures and unintended consequences

Pronovost PJ, Berenholtz SM, Needham DM. Translating evidence into practice: a model for large scale knowledge translation. BMJ. 2008 Oct 6;337.



bi·as

/ˈbīəs/

See definitions in:

All | Statistics | Bowls | Electronics | Needlework

noun

1. prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.
   "there was evidence of **bias against** foreign applicants"

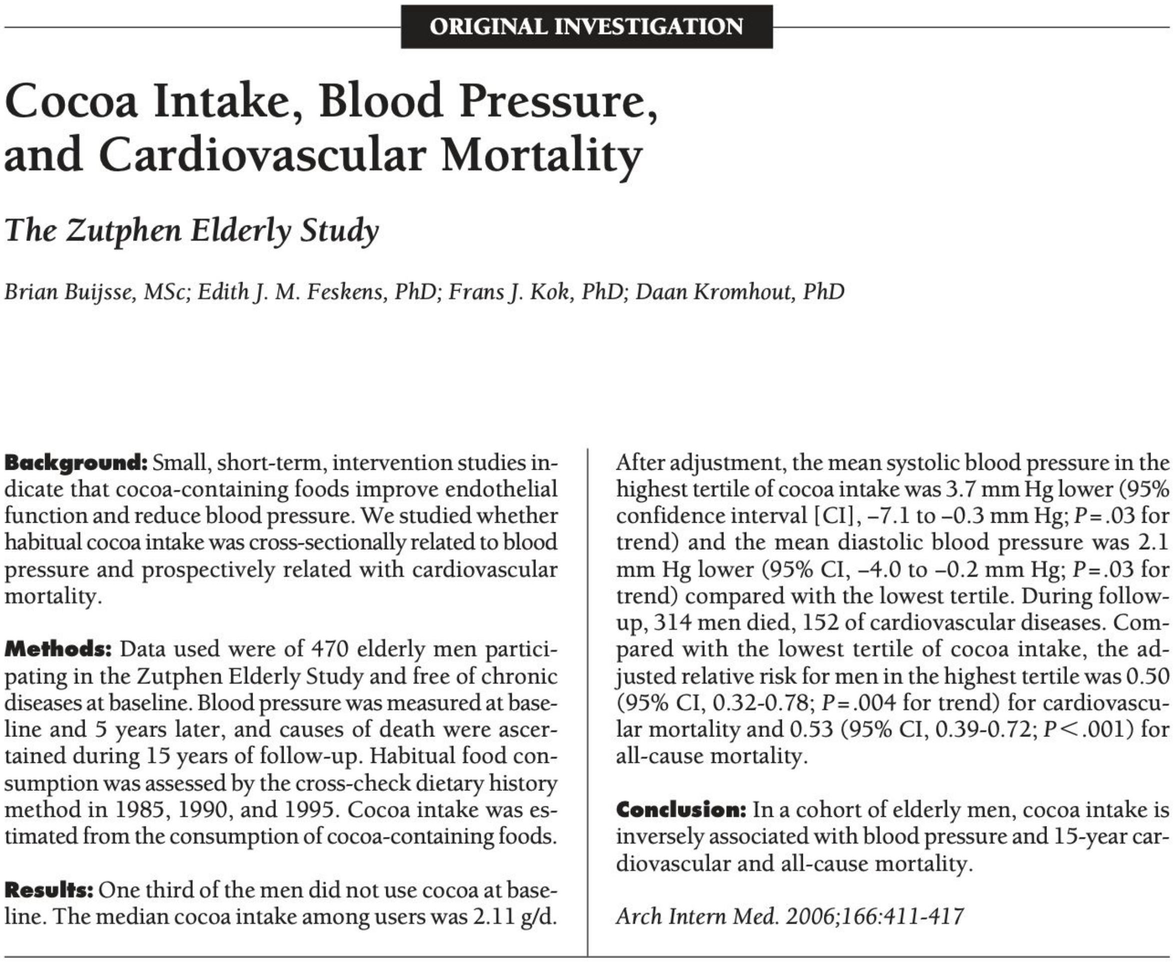
   Similar: prejudice | partiality | partisanship | favoritism | unfairness

Bias creates *distortions* in the translation between what is true and what we observe (usually using a model).

- A study reveals an association between a treatment and speed of recovery from a disease.

- A study reveals an association between a personal characteristic and mortality.

- A study reveals one or more risk factors for development of a particular disease.



ORIGINAL INVESTIGATION

Cocoa Intake, Blood Pressure, and Cardiovascular Mortality

The Zutphen Elderly Study

Brian Buijsse, MSc; Edith J. M. Feskens, PhD; Frans J. Kok, PhD; Daan Kromhout, PhD

Background: Small, short-term, intervention studies indicate that cocoa-containing foods improve endothelial function and reduce blood pressure. We studied whether habitual cocoa intake was cross-sectionally related to blood pressure and prospectively related with cardiovascular mortality.

Methods: Data used were of 470 elderly men participating in the Zutphen Elderly Study and free of chronic diseases at baseline. Blood pressure was measured at baseline and 5 years later, and causes of death were ascertained during 15 years of follow-up. Habitual food consumption was assessed by the cross-check dietary history method in 1985, 1990, and 1995. Cocoa intake was estimated from the consumption of cocoa-containing foods.

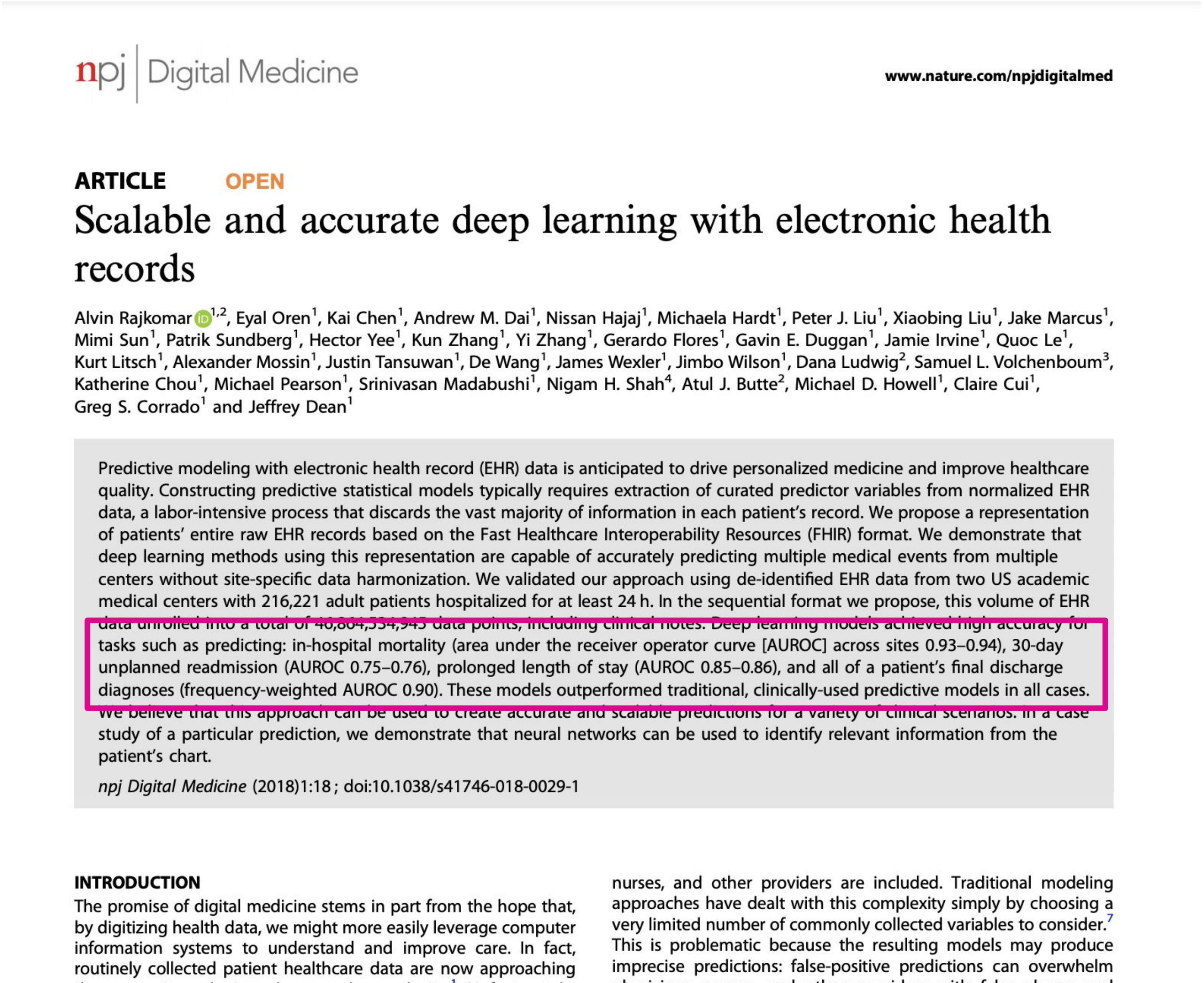
Results: One third of the men did not use cocoa at baseline. The median cocoa intake among users was 2.11 g/d. After adjustment, the mean systolic blood pressure in the highest tertile of cocoa intake was 3.7 mm Hg lower (95% confidence interval [CI], −7.1 to −0.3 mm Hg; *P*=.03 for trend) and the mean diastolic blood pressure was 2.1 mm Hg lower (95% CI, −4.0 to −0.2 mm Hg; *P*=.03 for trend) compared with the lowest tertile. During follow-up, 314 men died, 152 of cardiovascular diseases. Compared with the lowest tertile of cocoa intake, the adjusted relative risk for men in the highest tertile was 0.50 (95% CI, 0.32-0.78; *P*=.004 for trend) for cardiovascular mortality and 0.53 (95% CI, 0.39-0.72; *P*<.001) for all-cause mortality.

Conclusion: In a cohort of elderly men, cocoa intake is inversely associated with blood pressure and 15-year cardiovascular and all-cause mortality.

Arch Intern Med. 2006;166:411-417

Consider the following alternative explanations for the finding:

(1) fraud / mistakes
(2) random chance
(3) selection bias
(4) information bias
(5) confounding



npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two US academic medical centers with 216,221 adult patients hospitalized for at least 24 h. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient's final discharge diagnoses (frequency-weighted AUROC 0.90). These models outperformed traditional, clinically-used predictive models in all cases. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, we demonstrate that neural networks can be used to identify relevant information from the patient's chart.

npj Digital Medicine (2018)1:18 ; doi:10.1038/s41746-018-0029-1

INTRODUCTION

The promise of digital medicine stems in part from the hope that, by digitizing health data, we might more easily leverage computer information systems to understand and improve care. In fact, routinely collected patient healthcare data are now approaching

nurses, and other providers are included. Traditional modeling approaches have dealt with this complexity simply by choosing a very limited number of commonly collected variables to consider.[7] This is problematic because the resulting models may produce imprecise predictions: false-positive predictions can overwhelm

Consider the following alternative explanations for the finding:

(1) fraud / mistakes
(2) random chance
(3) selection bias
(4) information bias
(5) confounding

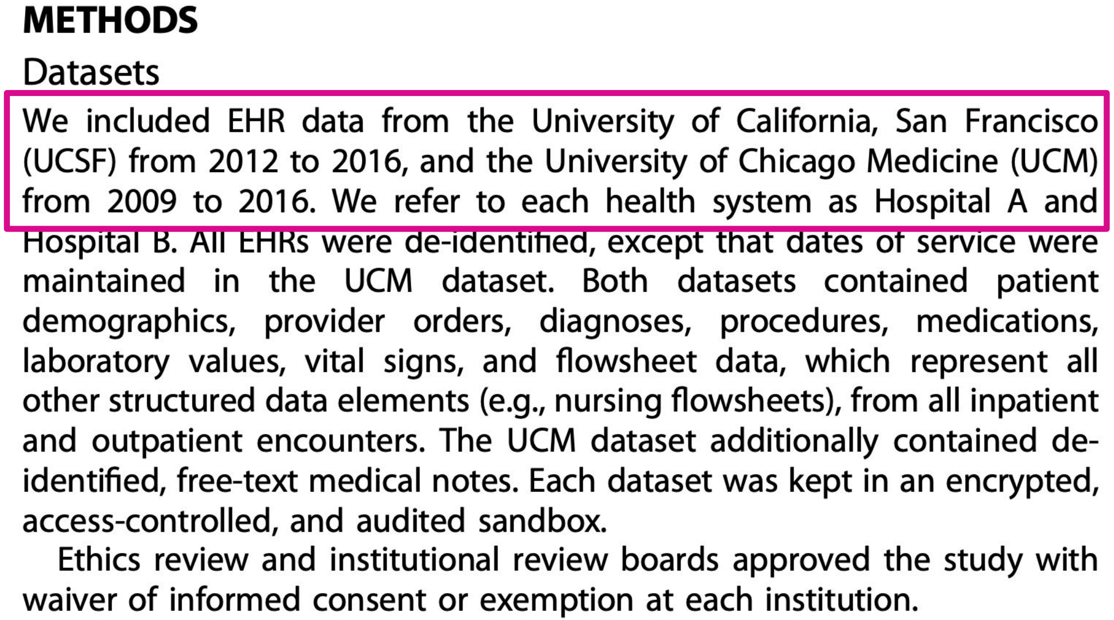
Machine learning in healthcare has raised issues of bias because:

- Machine learning models and data are too big for humans to manually address all biases (i.e. can’t just list them, stratify, etc.)
- Forces us to consider whether biases introduced by machine learning are qualitatively or quantitatively worse than biases in existing evidence
- Forces us to confront the issue of generalizability, and whether the way we treat patients does/should change in different contexts

Selection Bias

Selection bias: A form of bias that occurs when individuals or groups in a study differ systematically from the population of interest leading to a systematic error in an association or outcome.

Catalogue of Bias Collaboration, Nunan D, Bankhead C, Aronson JK. Selection bias. Catalogue Of Bias 2017:
http://www.catalogofbias.org/biases/selection-bias/



METHODS

Datasets

We included EHR data from the University of California, San Francisco (UCSF) from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016. We refer to each health system as Hospital A and Hospital B. All EHRs were de-identified, except that dates of service were maintained in the UCM dataset. Both datasets contained patient demographics, provider orders, diagnoses, procedures, medications, laboratory values, vital signs, and flowsheet data, which represent all other structured data elements (e.g., nursing flowsheets), from all inpatient and outpatient encounters. The UCM dataset additionally contained de-identified, free-text medical notes. Each dataset was kept in an encrypted, access-controlled, and audited sandbox.

Ethics review and institutional review boards approved the study with waiver of informed consent or exemption at each institution.

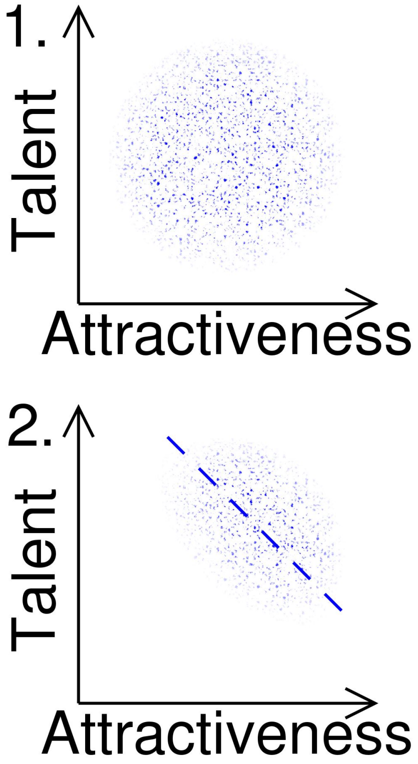
Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine. 2018 May 8;1(1):1-0.

Examples of Selection Bias:

- **Ascertainment Bias / Sampling Bias:** Systematic differences in the identification of individuals included in a study or distortion in the collection of data in a study.

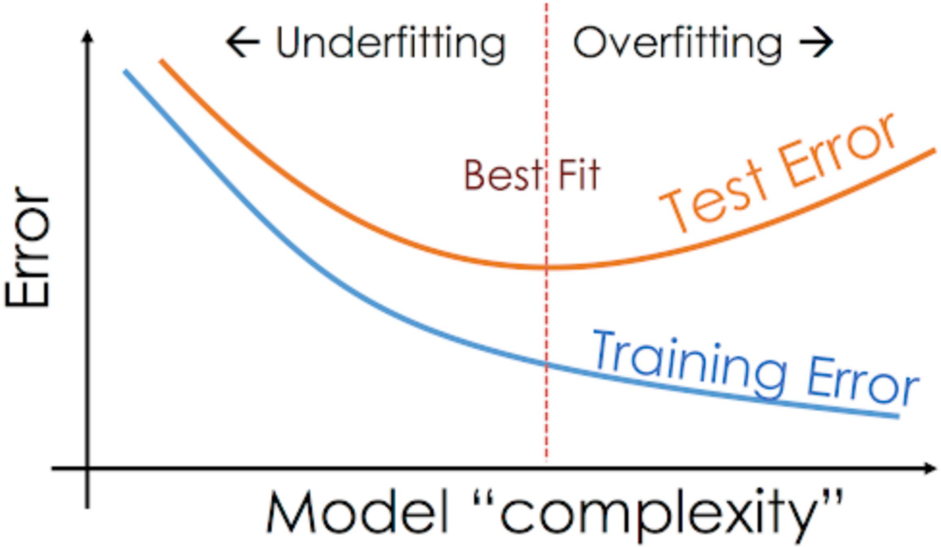
- **Informed Presence Bias:** The presence of a person’s information in an electronic health record is affected by the person’s health status and other factors; amount of recorded data is informative.

- **Berkson’s Bias / Collider Bias:** Correlations appear in subgroups of overall population even though two variables are independent in the overall population.



By Cmglee - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=91398290

Overfitting Viewed as Selection Bias



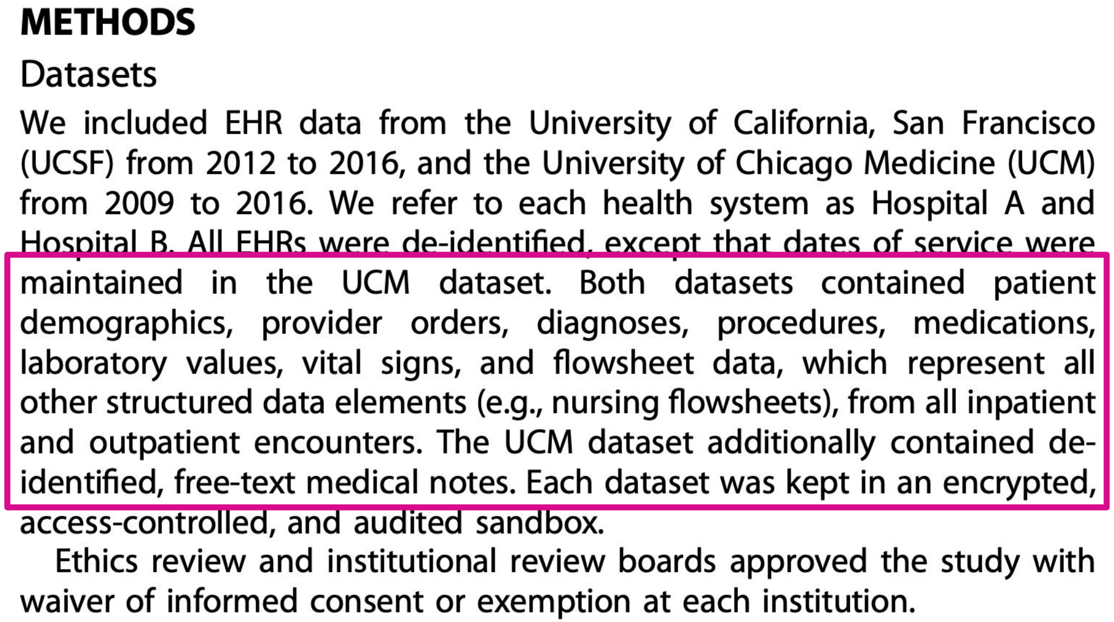
https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/

Information Bias

Information bias: Bias that arises from systematic differences in the collection, recall, recording or handling of information used in a study.

In EHR studies, information bias is often *the cause of* selection bias, since you’re using previously-recorded information to generate cohorts.

Catalogue of bias collaboration. Bankhead CR, Spencer EA, Nunan D. Information bias. In: Sackett Catalogue Of Biases 2019. https://catalogofbias.org/biases/information-bias/



METHODS

Datasets

We included EHR data from the University of California, San Francisco (UCSF) from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016. We refer to each health system as Hospital A and Hospital B. All EHRs were de-identified, except that dates of service were maintained in the UCM dataset. Both datasets contained patient demographics, provider orders, diagnoses, procedures, medications, laboratory values, vital signs, and flowsheet data, which represent all other structured data elements (e.g., nursing flowsheets), from all inpatient and outpatient encounters. The UCM dataset additionally contained de-identified, free-text medical notes. Each dataset was kept in an encrypted, access-controlled, and audited sandbox.

Ethics review and institutional review boards approved the study with waiver of informed consent or exemption at each institution.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine. 2018 May 8;1(1):1-0.

Examples of Information Bias:

- **Misclassification / Measurement Bias:** Incorrectness or absence of measurement is correlated with factors relevant to the study. Sometimes used interchangeably with “information bias”.

- **Observer Bias:** Systematic difference between truth and recorded data due to observer variation. Especially bad when observer properties correlated with other aspects of the population under study.

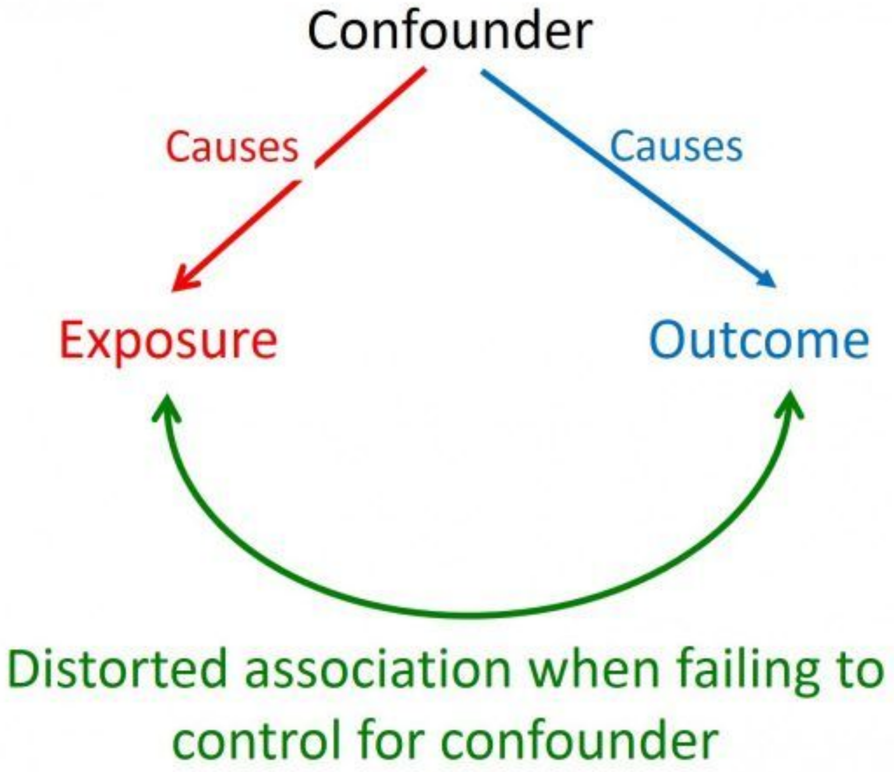
- **Recall Bias:** Participants do not remember previous events or experiences accurately or omit details. Self-reported info in EHRs strongly impacted.

- **Reporting Bias:** Normally used in discussions of scientific misconduct, this is when there is selective disclosure or withholding of information based on the nature of that information.

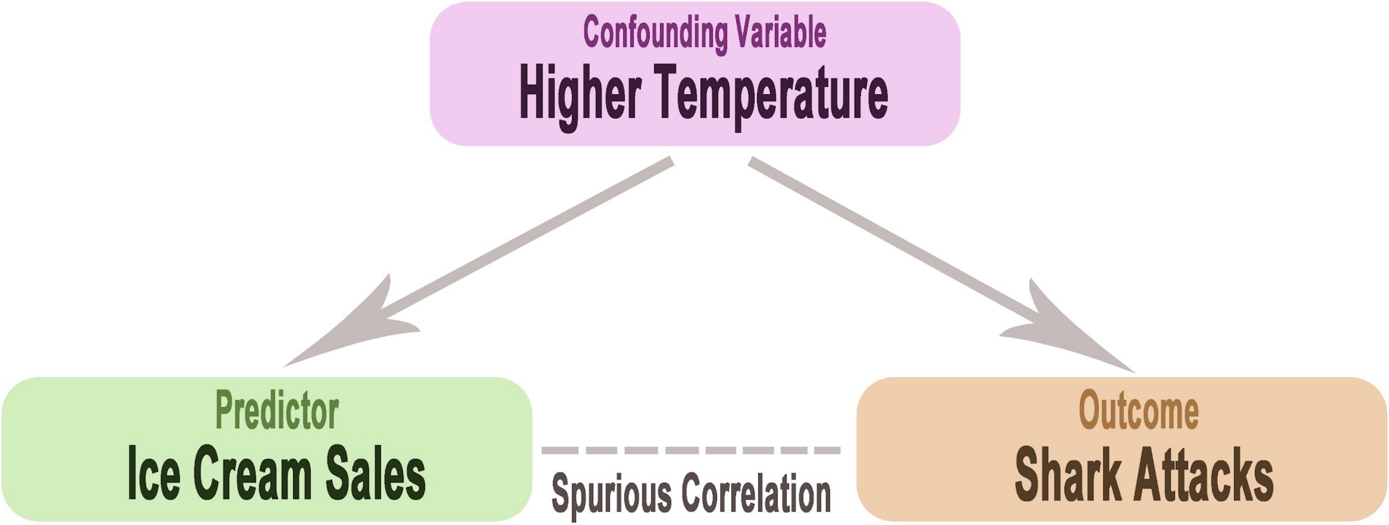
Confounding

Confounding: A distortion that modifies an association between an exposure and an outcome because a factor is independently associated with the exposure and the outcome.

Catalogue of bias collaboration, Aronson JK, Bankhead C, Nunan D. Confounding. In Catalogue Of Biases. 2018. www.catalogueofbiases.org/biases/confounding


Confounder
Causes
Causes
Exposure
Outcome
Distorted association when failing to control for confounder

Catalogue of bias collaboration, Aronson JK, Bankhead C, Nunan D. Confounding. In Catalogue Of Biases. 2018. www.catalogueofbiases.org/biases/confounding


Confounding Variable

Higher Temperature

Predictor

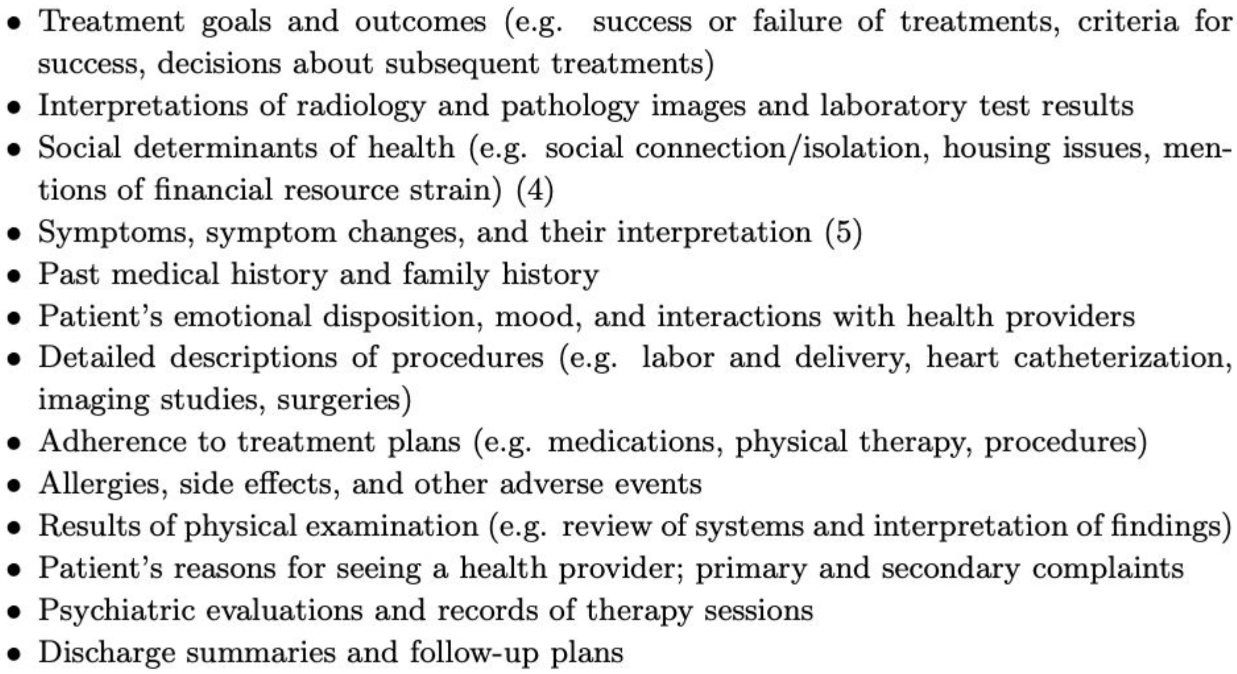
Ice Cream Sales

Spurious Correlation

Outcome

Shark Attacks

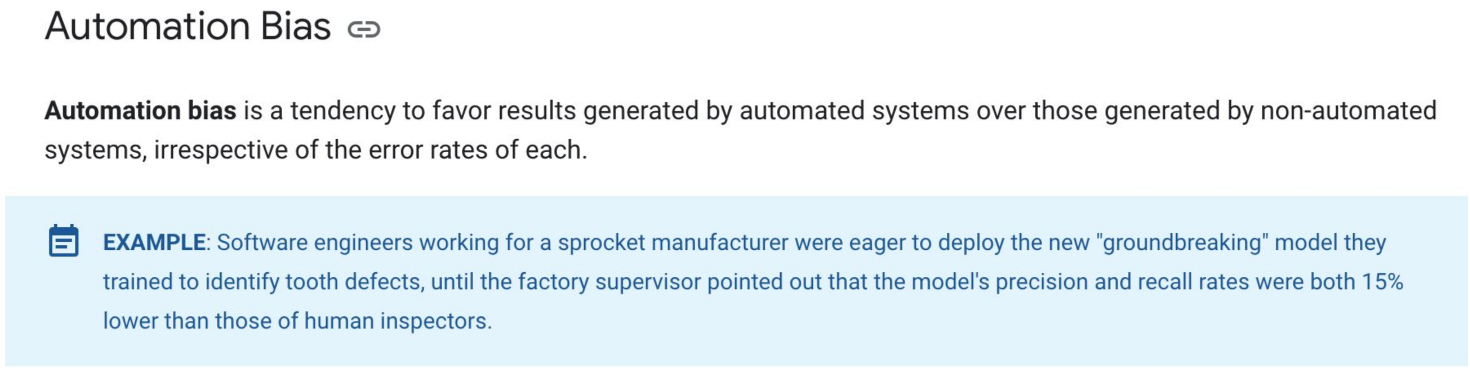
Unknown confounders in machine learning studies



- Treatment goals and outcomes (e.g. success or failure of treatments, criteria for success, decisions about subsequent treatments)
- Interpretations of radiology and pathology images and laboratory test results
- Social determinants of health (e.g. social connection/isolation, housing issues, mentions of financial resource strain) (4)
- Symptoms, symptom changes, and their interpretation (5)
- Past medical history and family history
- Patient's emotional disposition, mood, and interactions with health providers
- Detailed descriptions of procedures (e.g. labor and delivery, heart catheterization, imaging studies, surgeries)
- Adherence to treatment plans (e.g. medications, physical therapy, procedures)
- Allergies, side effects, and other adverse events
- Results of physical examination (e.g. review of systems and interpretation of findings)
- Patient's reasons for seeing a health provider; primary and secondary complaints
- Psychiatric evaluations and records of therapy sessions
- Discharge summaries and follow-up plans

This list is from my recent article “Modern Clinical Text Mining” - it’s a list of the fields most likely to be found in text, not a list of unknown confounders. But the two are correlated!

Other Human Biases

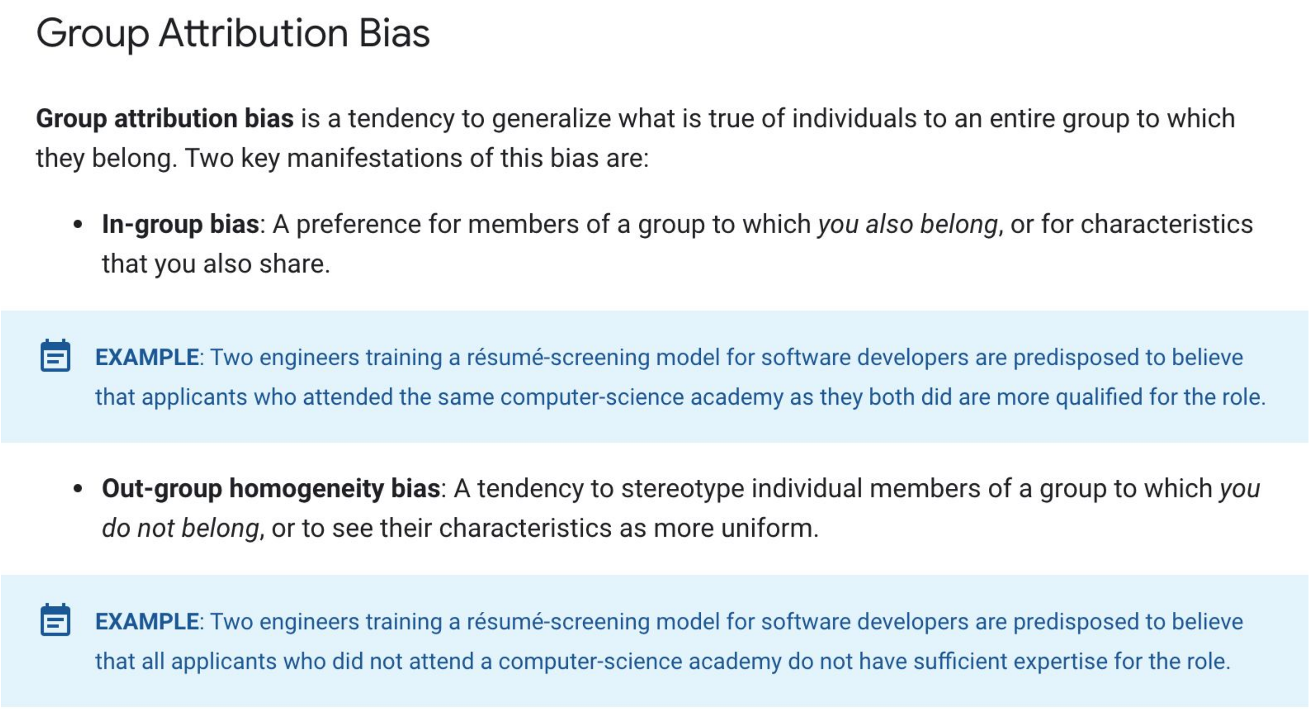


Automation Bias

Automation bias is a tendency to favor results generated by automated systems over those generated by non-automated systems, irrespective of the error rates of each.

EXAMPLE: Software engineers working for a sprocket manufacturer were eager to deploy the new "groundbreaking" model they trained to identify tooth defects, until the factory supervisor pointed out that the model's precision and recall rates were both 15% lower than those of human inspectors.

https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias



Group Attribution Bias

Group attribution bias is a tendency to generalize what is true of individuals to an entire group to which they belong. Two key manifestations of this bias are:

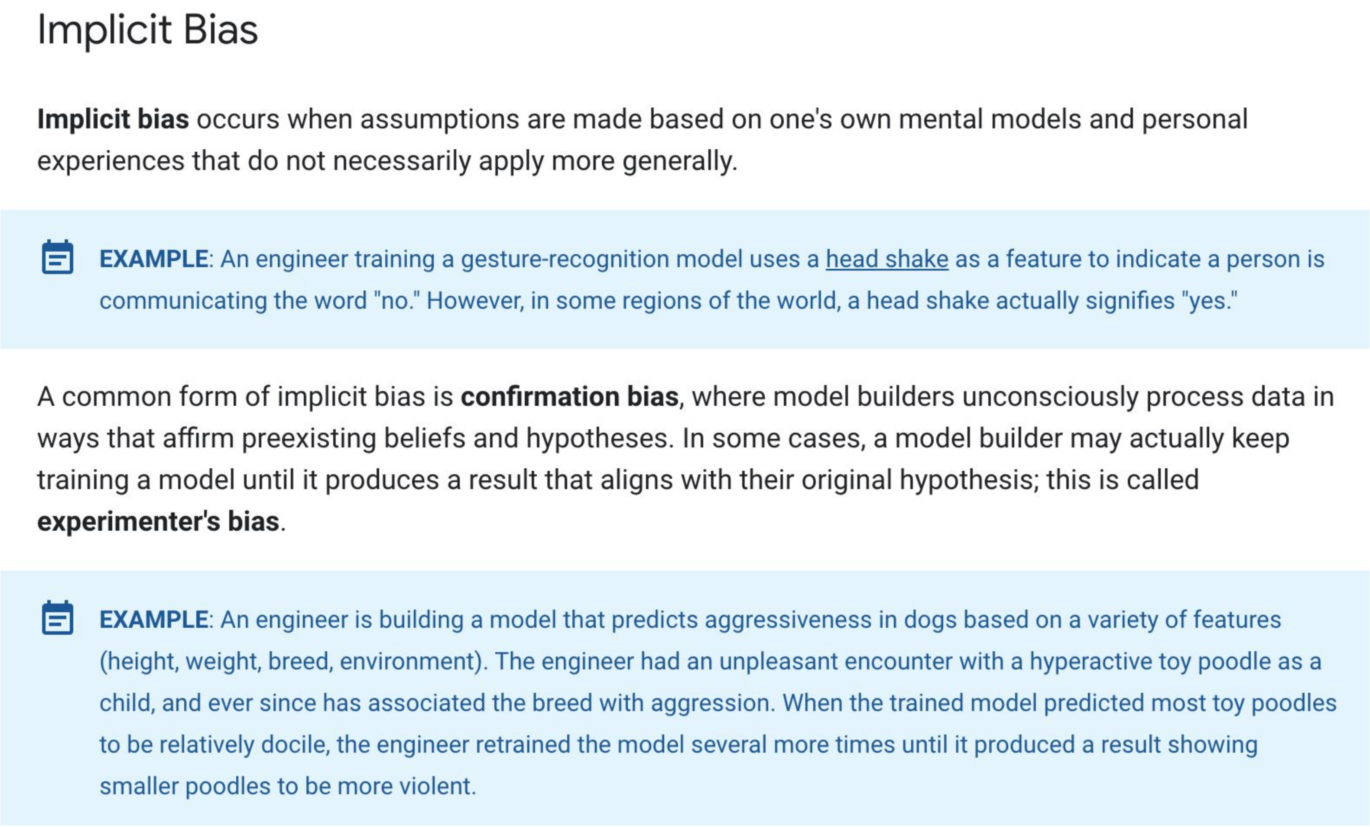
- **In-group bias**: A preference for members of a group to which *you also belong*, or for characteristics that you also share.

EXAMPLE: Two engineers training a résumé-screening model for software developers are predisposed to believe that applicants who attended the same computer-science academy as they both did are more qualified for the role.

- **Out-group homogeneity bias**: A tendency to stereotype individual members of a group to which *you do not belong*, or to see their characteristics as more uniform.

EXAMPLE: Two engineers training a résumé-screening model for software developers are predisposed to believe that all applicants who did not attend a computer-science academy do not have sufficient expertise for the role.

https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias



Implicit Bias

Implicit bias occurs when assumptions are made based on one's own mental models and personal experiences that do not necessarily apply more generally.

EXAMPLE: An engineer training a gesture-recognition model uses a head shake as a feature to indicate a person is communicating the word "no." However, in some regions of the world, a head shake actually signifies "yes."

A common form of implicit bias is **confirmation bias**, where model builders unconsciously process data in ways that affirm preexisting beliefs and hypotheses. In some cases, a model builder may actually keep training a model until it produces a result that aligns with their original hypothesis; this is called **experimenter's bias**.

EXAMPLE: An engineer is building a model that predicts aggressiveness in dogs based on a variety of features (height, weight, breed, environment). The engineer had an unpleasant encounter with a hyperactive toy poodle as a child, and ever since has associated the breed with aggression. When the trained model predicted most toy poodles to be relatively docile, the engineer retrained the model several more times until it produced a result showing smaller poodles to be more violent.

https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

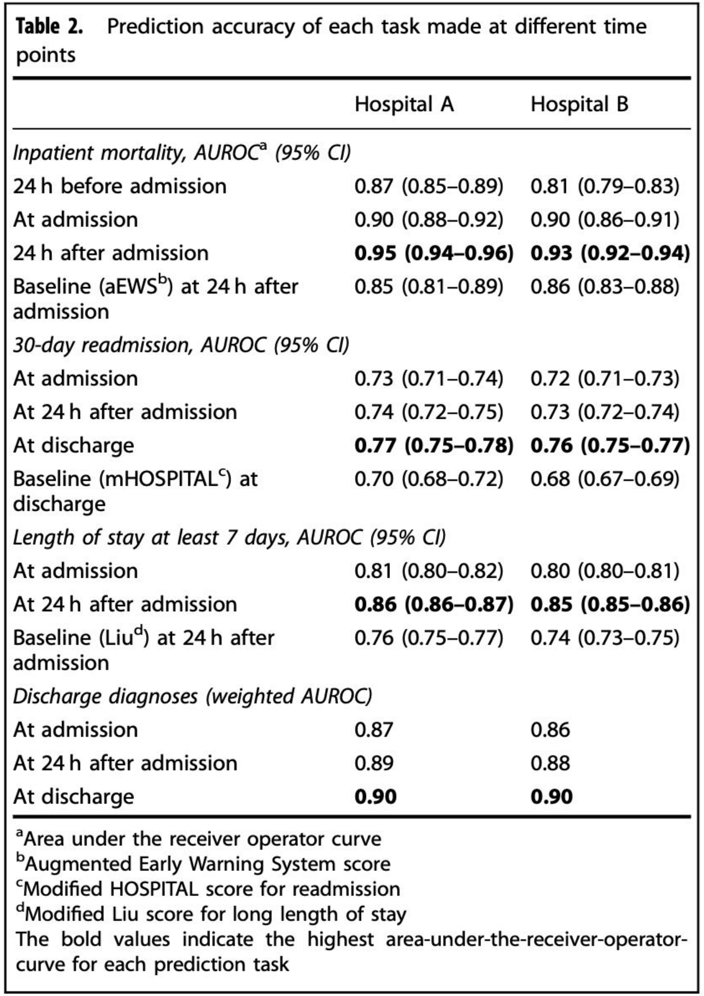


Table 2. Prediction accuracy of each task made at different time points

	Hospital A	Hospital B
Inpatient mortality, AUROC[a] (95% CI)		
24 h before admission	0.87 (0.85–0.89)	0.81 (0.79–0.83)
At admission	0.90 (0.88–0.92)	0.90 (0.86–0.91)
24 h after admission	**0.95 (0.94–0.96)**	**0.93 (0.92–0.94)**
Baseline (aEWS[b]) at 24 h after admission	0.85 (0.81–0.89)	0.86 (0.83–0.88)
30-day readmission, AUROC (95% CI)		
At admission	0.73 (0.71–0.74)	0.72 (0.71–0.73)
At 24 h after admission	0.74 (0.72–0.75)	0.73 (0.72–0.74)
At discharge	**0.77 (0.75–0.78)**	**0.76 (0.75–0.77)**
Baseline (mHOSPITAL[c]) at discharge	0.70 (0.68–0.72)	0.68 (0.67–0.69)
Length of stay at least 7 days, AUROC (95% CI)		
At admission	0.81 (0.80–0.82)	0.80 (0.80–0.81)
At 24 h after admission	**0.86 (0.86–0.87)**	**0.85 (0.85–0.86)**
Baseline (Liu[d]) at 24 h after admission	0.76 (0.75–0.77)	0.74 (0.73–0.75)
Discharge diagnoses (weighted AUROC)		
At admission	0.87	0.86
At 24 h after admission	0.89	0.88
At discharge	**0.90**	**0.90**

[a] Area under the receiver operator curve
[b] Augmented Early Warning System score
[c] Modified HOSPITAL score for readmission
[d] Modified Liu score for long length of stay
The bold values indicate the highest area-under-the-receiver-operator-curve for each prediction task

Some thoughts:

- Reporting of improvement over baseline models
- Reporting of performance on discharge diagnoses as an average over 10,000+ codes
- Reporting of deep learning models only vs. simpler learned models
- Inclusion of covariates more likely to be associated with physician decision making rather than patient disease state (although this could be true of the whole EHR)

The big question:

Are biases in machine learning studies of EHR data enough to make it untrustworthy?

Do current efforts at producing generalizable models (e.g., federated learning, common data models) adequately address this issue?