SMU Singapore Management University
School of Computing and Information Systems

2021-22 Term 1

IS457: Fairness in Socio-technical Systems

Week 2 - Case studies of measuring fairness and bias (I)

KWAK Haewoon

Image from Flaticon.com

Study questions

What are the potential contributions of AI in healthcare?

What is a desirable bias and undesirable bias in healthcare?

How different are the origin of the biases between experimental/clinical data and digital biomarkers data?

Why did the high-risk care management algorithm show racial biases?

What biases can exist in a model with the same performance in a criminal justice system?

What is over-policing and how is it reinforced?

Promise of AI in healthcare

Big data and machine learning can help healthcare providers by

- Advancing medical knowledge
- Automating the routine
- Democratizing expertise

Advancing medical knowledge

AI analyzes data about how humans work and how to care for them for

- improve care
- discovering new treatments
- advancing scientific hypotheses

Examples: Personalized treatment, discovering drug, prediction the onset of illness, predicting patients' reaction to particular treatment, and more.

Price, I. I., and W. Nicholson. "Medical AI and contextual bias." (2019).

AI in discovering drugs

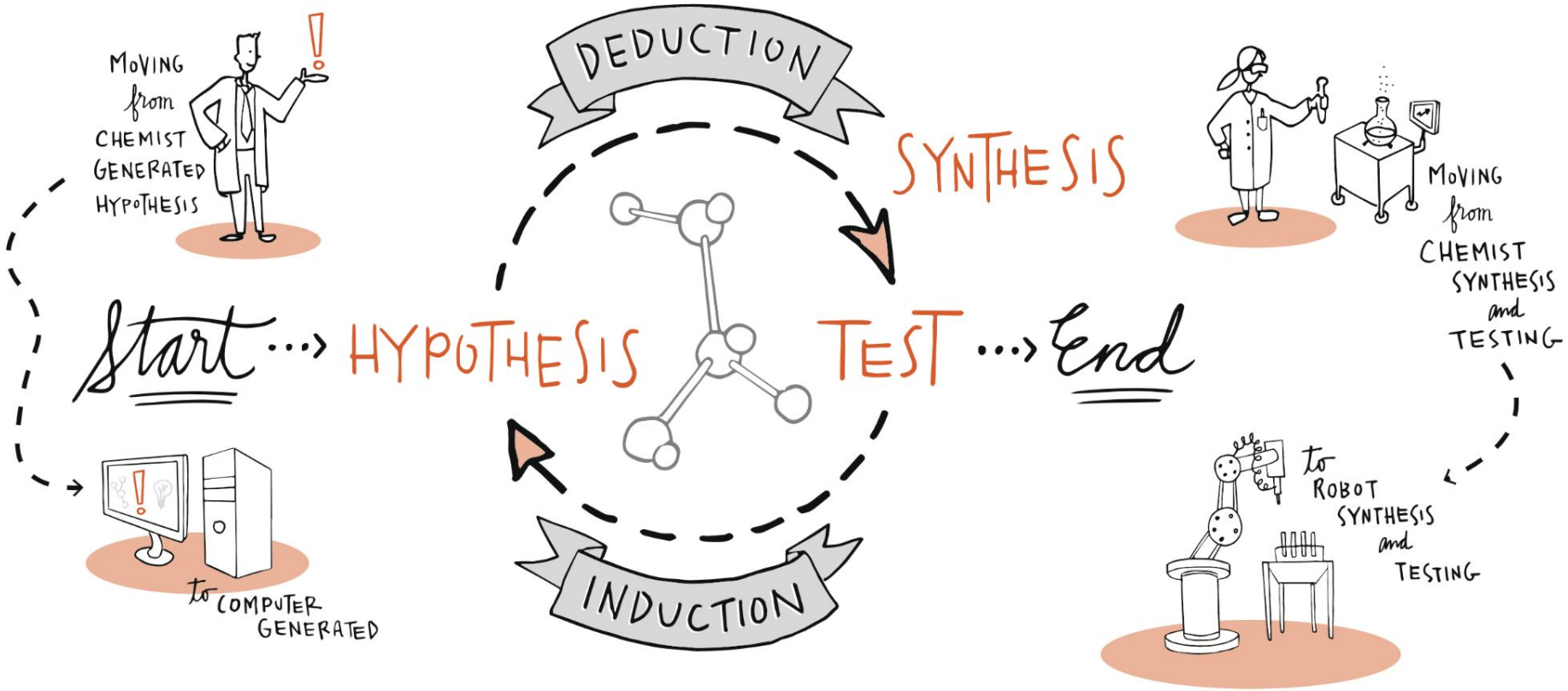


Image from Schneider, Gisbert. "Mind and machine in drug design." Nature Machine Intelligence 1.3 (2019): 128-130.

Automating the routine (I)

Much of medical practice consists of paperwork and routine tasks that often don’t do much to help patients and contributes to physician burnout.

Image from https://www.unmc.edu/news.cfm?match=22277

Automating the routine (II)

Physicians spent almost half of their time on electronic health record work and desk work, and only a quarter of their time seeing patients.

Even in the examination room, physicians spent only about half of their time interacting with patients — and about a third interacting with electronic health records and desk work.

Sinsky, Christine, et al. "Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties." Annals of internal medicine 165.11 (2016): 753-760.

Automating the routine (II)

AI could improve medical practice by:

- identifying and highlighting the most relevant medical information from patient medical records.
- providing the most relevant medical literature to doctors based on natural-language processing.
- transcribing patient conversations and provider notes.

Democratizing expertise (I)

There are tremendous differences in the quality and level of care patients receive based on the context in which they receive that care.

Image from https://emansion.gov.lr/2press.php?news_id=994&related=7&pg=sp
https://hub.jhu.edu/2013/07/16/us-news-hospital-rankings/

Liberia, as of 2016, there were 298 doctors for a population of 4.5M, including only 15 pediatricians and 6 ophthalmologists.

In rural India, a single doctor can be responsible for as many as 30k residents.

Democratizing expertise (II)

AI promises to reduce this variation by providing care at the level of excellent specialists.

- Diagnostics
- Treatment recommendations

Diagnostics

Diagnosis is the process of figuring out what's wrong with a patient.

Diagnosis depends on recognizing the right symptoms and using them to identify underlying problems from a vast realm of possibilities.

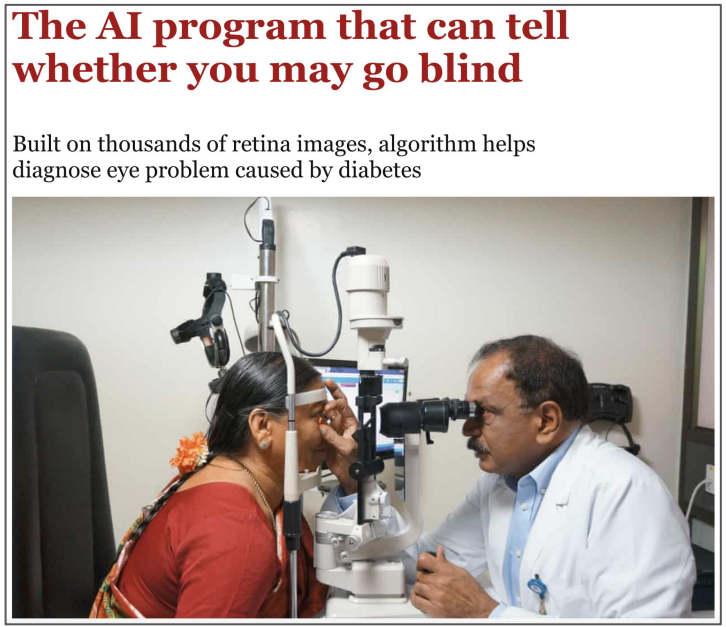
Providers may reach incorrect diagnoses because:

- They never acquired the relevant medical knowledge.
- The knowledge they acquired is outdated.
- They lack time to conduct the relevant research.
- They suffer from heuristic biases.

EyeDiagnosis' IDx-DR

IDx-DR

IDx-DR is approved by FDA for autonomous diagnosis and performs at a level comparable to ophthalmologists, even when operated by novices.



The AI program that can tell whether you may go blind

Built on thousands of retina images, algorithm helps diagnose eye problem caused by diabetes

Similar technology is also implemented by Google AI team and local experts in India

Treatment recommendations

Once providers have determined what ails the patient, they must select from a menu of possibilities to determine the best option for improvement.

E.g., knowing that a patient has a certain type of cancer, a well-trained and experienced oncologist knows that the best course of treatment is:

- Surgery, radiotherapy, chemotherapy, or some combination
- Which drugs or protocols are likely most effective

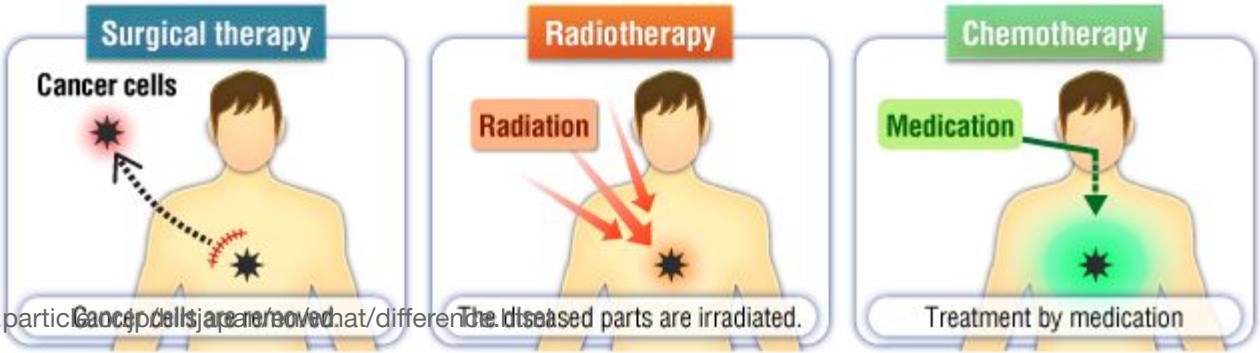
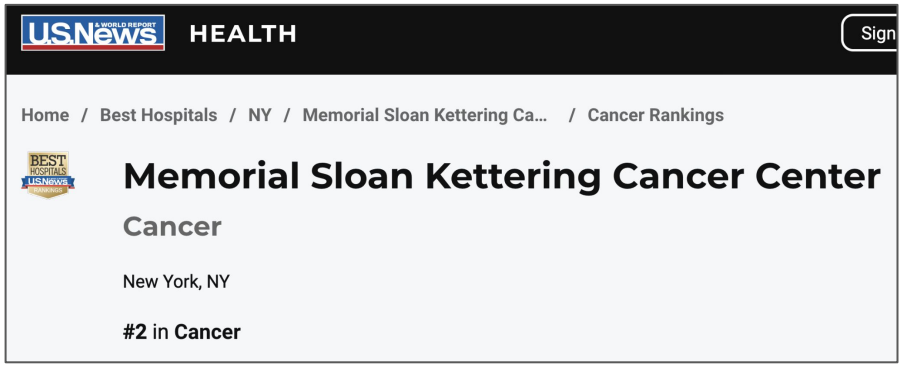


Image from https://www.particlecancercare.com/what/different-treatments

AI in treatment recommendation

E.g., IBM's Watson Oncology

- uses machine-learning-based natural language processing to analyze patient medical records to determine cancer type.
- recommends treatment based on what oncologists at Memorial Sloan Kettering would do when faced with a similar patient.



https://health.usnews.com/best-hospitals/area/ny/memorial-sloan-kettering-cancer-center-6213060/cancer

Watson Oncology to democratize medical expertise

IBM licenses Watson Oncology for use at hundreds of hospitals worldwide and has evaluated its performance at hospitals in Thailand, India, and Mexico.

> “Oncologists . . . felt [Watson Oncology] would be particularly beneficial in clinics that lack subspecialist expertise.”

(Of course, some limitations/risks have been reported as well)

npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare

Davide Cirillo[1,10]✉, Silvina Catuara-Solarz[2,3,10], Czuee Morey[3,4], Emre Guney[5], Laia Subirats[6,7], Simona Mellino[3], Annalisa Gigante[3], Alfonso Valencia[1,8], María José Rementeria[1], Antonella Santuccione Chadha[3] and Nikolaos Mavridis[3,9]

https://www.nature.com/articles/s41746-020-0288-5

Desirable bias and undesirable bias

Desirable bias

A desirable bias implies taking into account sex and gender differences to make a precise diagnosis and recommend a tailored and more effective treatment for each individual.

Undesirable bias

An undesirable bias is that which exhibits unintended or unnecessary sex and gender discrimination.

Examples of undesirable biases

Misrepresentation of the target population — insufficient representation of pregnant women in psychiatric research.

Depression is frequently observed among women. However, this may result from a skewed diagnosis due to clinical scales of depression measuring symptoms (e.g., crying, sadness) that occur more frequently among women.

Male-type depression symptoms scale (MDS)

Item Description	% (SE)		
	Total	Men	Women
Stress	68.9 (1.6)	63.3 (1.9)	75.2 (2.4)[a]
Irritability	90.3 (1.4)	86.6 (1.9)	94.7 (1.9)[a]
Anger attacks/aggression	92.05 (1.2)	94.85 (1.9)[a]	88.94 (1.4)
Sleep problems	37.7 (1.5)	29.2 (2.1)	47.1 (1.9)[a]
Alcohol/other drug abuse	51.6 (1.9)	61.4 (3.0)[a]	40.6 (1.9)
Loss of interest	89.7 (.973)	87.8 (1.5)	91.8 (1.0)[b]
Risk-taking behavior	41.6 (1.5)	52.7 (2.0)[a]	29.1 (2)
Hyperactivity	57.9 (1.4)	57.6 (92.1)	58.4 (1.8)
Mean score	6.06	6.05	6.07
Prevalence, %	23.8	26.3[c]	21.9

[a] $P \leq .001$.
[b] $P \leq .05$.
[c] $P \leq .01$.

Martin, Lisa A., Harold W. Neighbors, and Derek M. Griffith. "The experience of symptoms of depression in men vs women: analysis of the National Comorbidity Survey Replication." JAMA psychiatry 70.10 (2013): 1100-1106.

Sources and types of health data

Experimental and clinical data

Digital biomarkers

Image from https://en.wikipedia.org/wiki/Animal_testing_on_rodents

https://news.samsung.com/global/electrocardiogram-monitoring-cleared-for-galaxy-watch-active2-by-south-koreas-ministry-of-food-and-drug-safety

More males in experimental and clinical data

In early 2000, sex-specific biological differences were neglected.

Both experimental and clinical studies focused on male experimental models or male subjects.

Even nowadays, male mouse models are overall more represented than female models in basic, preclinical, and surgical biomedical research.

Origin and impact of those biases

The lack of representation of female models and patients is partly due to technical and bioethical considerations:

- Reduce the impact of estrous cycle in experimental studies
- Protect women of childbearing age in clinical research.

Some of the treatments that currently exist for several diseases are not adequately evaluated in women, who are likely to be underrepresented in clinical trials, especially in Phases I and II.

Example - Zolpidem (2013)

Zolpidem, a widely prescribed insomnia drug

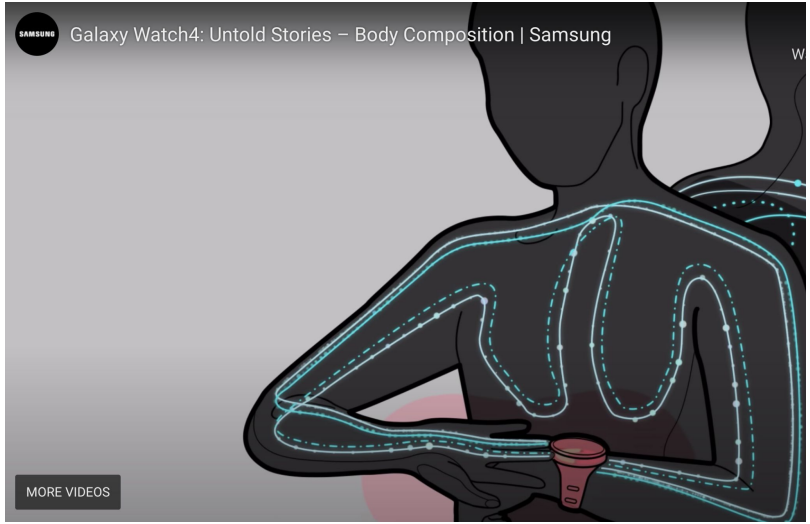
“Women appear to be more susceptible to this risk [of next-morning impairment] because they eliminate zolpidem from their bodies more slowly than men.”

“FDA has informed the manufacturers that the recommended dose of zolpidem **for women** should be lowered **from 10 mg to 5 mg** for immediate-release products (Ambien, Edluar, and Zolpimist) and **from 12.5 mg to 6.25 mg** for extended-release products (Ambien CR)”

https://www.fda.gov/files/drugs/published/Drug-Safety-Communication--Risk-of-next-morning-impairment-after-use-of-insomnia-drugs--FDA-requires-lower-recommended-doses-for-certain-drugs-containing-zolpidem-%28Ambien--Ambien-CR--Edluar--and-Zolpimist%29.pdf

Digital biomarkers

Physiological, psychological and behavioral indicators based on data including human-computer interaction, physical activity, and voice variations, collected by portable, wearable, implantable or even ingestible devices.



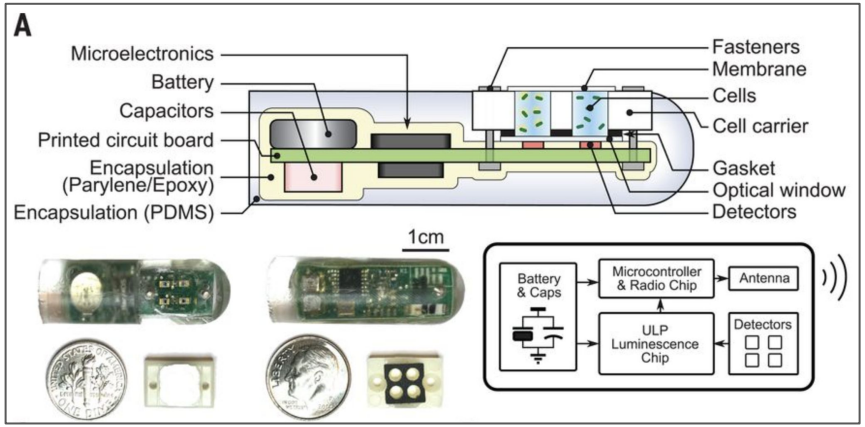
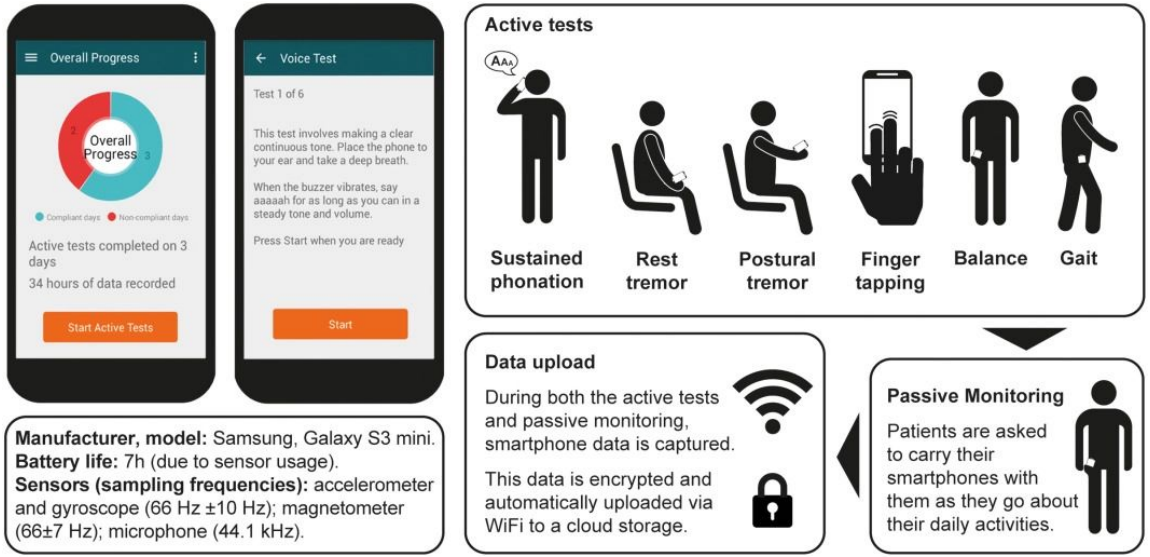


Image from https://www.youtube.com/watch?v=7sFa6HWkseU

Mimee, Mark, et al. "An ingestible bacterial-electronic system to monitor gastrointestinal health." Science 360.6391 (2018): 915-918.

Biases in digital biomarker datasets

In a study assessing digital biomarkers for Parkinson's disease (PD), only 18.6% were women.



Lipsmeier, Florian, et al. "Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial." Movement Disorders 33.8 (2018): 1287-1297.

Biases from digital devices

Sao2 (arterial hemoglobin oxygen saturation) level, sensor type, **skin color**, and **gender** were predictive of errors in Spo2 (pulse oximetry) estimates at low Sao2 levels.

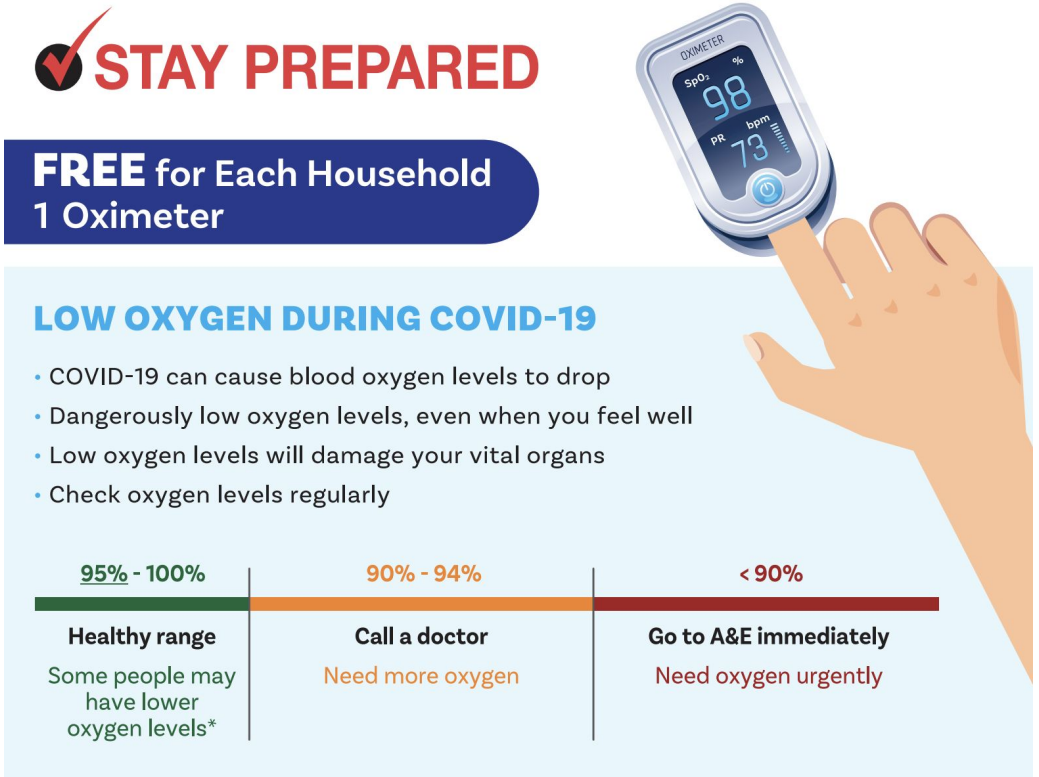
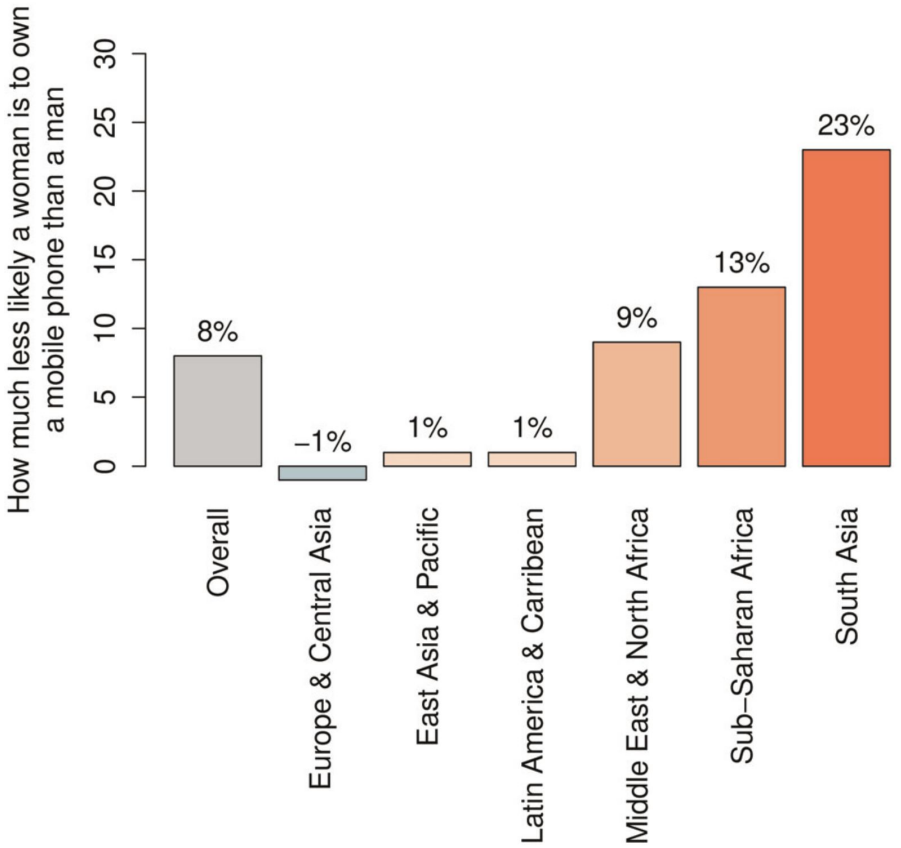


Image from https://stayprepared.sg/oximeter/media-library/

Feiner, John R., John W. Severinghaus, and Philip E. Bickler. "Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender." Anesthesia & Analgesia 105.6 (2007): S18-S23.

Biases from unequal access and use of digital devices

In low and middle income countries (“overall” in the graph), women are 8% less likely to own smartphones and 26% less likely to use the internet compared with men.



Lab: Google Teachable Machine

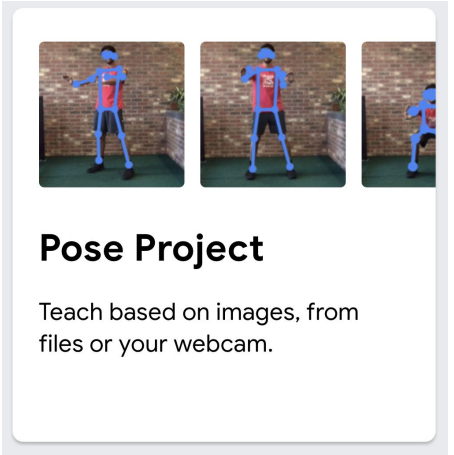
Individual activity (15 mins) ~ 9:40

Follow the tutorial:

https://medium.com/@warronbebster/teachable-machine-tutorial-head-tilt-f4f6116f491

Now you know how to build a classification model!

Build your own classification for whatever you are interested in.



Group activity: Experiment 1 (10:00am~)

Download following datasets:

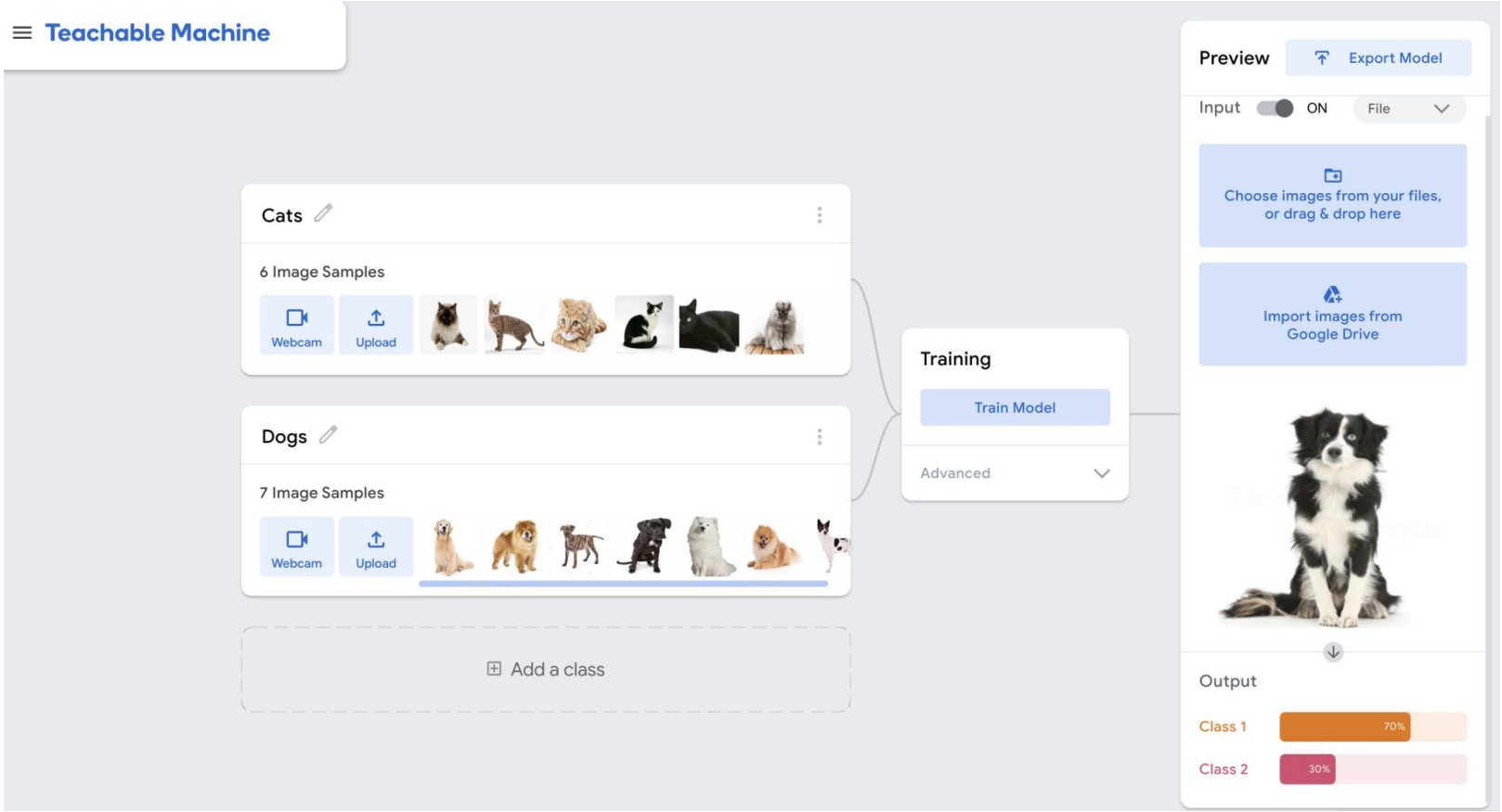
1. Cat-training-dataset-12.zip (for a “Cat” class)
2. Dog-training-dataset-7.zip (for a “Dog” class)
3. Testing-dataset.zip (for testing)

Build a classification model for cats and dogs by using dataset 1 and 2.

Test your model by using dataset 3.

Fill the spreadsheet “Experiment 1” Tab:
https://docs.google.com/spreadsheets/d/16kBgzNn-sNcLYN3-b0XElP0cCDbQsfq2ktDaEHevAo/edit?usp=sharing

Your screen should look like this...



Group activity: Experiment 2

Download following datasets:

1. Cat-training-dataset-6.zip (for a “Cat” class)
2. Dog-training-dataset-7.zip (for a “Dog” class)
3. Testing-dataset.zip (for testing)

Build a **NEW** classification model for cats and dogs by using dataset 1 and 2.

Test your model by using dataset 3.

Fill the spreadsheet “Experiment 2” Tab:
https://docs.google.com/spreadsheets/d/16kBgzNn-sNcLYN3-b0XElP0cCDbQsfq2ktDaEHevAo/edit?usp=sharing

Group activity: Experiment 3

Download following datasets:

1. Cat-training-dataset-6.zip (for a “Cat” class)
2. Dog-revised-training-dataset-6.zip (for a “Dog” class)
3. Testing-dataset.zip (for testing)

Build a **NEW** classification model for cats and dogs by using dataset 1 and 2.

Test your model by using dataset 3.

Fill the spreadsheet “Experiment 3” Tab:
https://docs.google.com/spreadsheets/d/16kBgzNn-sNcLYN3-b0XElP0cCDbQsfq2ktDaEHevAo/edit?usp=sharing

Where is medical AI (typically) trained?

IT companies & Academic medical centers or state-of-the-art hospitals

- IBM's Watson Oncology & Memorial Sloan Kettering.
- EyeDiagnosis & Univ. of Iowa Health System and the Univ. of Arizona.
- MIMIC dataset & Beth Israel Deaconess Medical Center, a high-resource Harvard-affiliated hospital in Boston.

Why?

Data

Reputation

Legal considerations

Group activity: Bias in AI healthcare ~11:17am

Download electronic health records in two different hospitals in an alien planet.

AI for diagnosis and treatment recommendation is trained in the high-resource hospital and is applied to the low-resource hospitals. What will be the potential risks here?

(Hint: Examine the data from the perspectives of different population, different disease, different resources, and different cost.)

After the group discussion, please share your thoughts on Padlet!

https://padlet.com/haewoon/IS457

RESEARCH

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366.6464 (2019): 447-453.

“High-risk care management” programs

“High-risk care management” programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers.

Large health systems rely on the algorithm to target patients for these programs.

- One of the largest commercial risk-prediction tools that are applied to roughly 200M people in the U.S each year.

Prioritization of patients is unavoidable

As high-risk care management is expensive (resource-wise),

- costs going toward teams of dedicated nurses
- extra primary care appointment slots
- other scarce resources

Health systems rely on algorithms to identify patients who will benefit the most.

“Better, earlier access to care for the high-risk patients will prevent burdensome and costly complications, thereby achieving both higher quality and lower cost.”

Bates, David W., et al. "Big data in health care: using analytics to identify and manage high-risk and high-cost patients." Health affairs 33.7 (2014): 1123-1131.

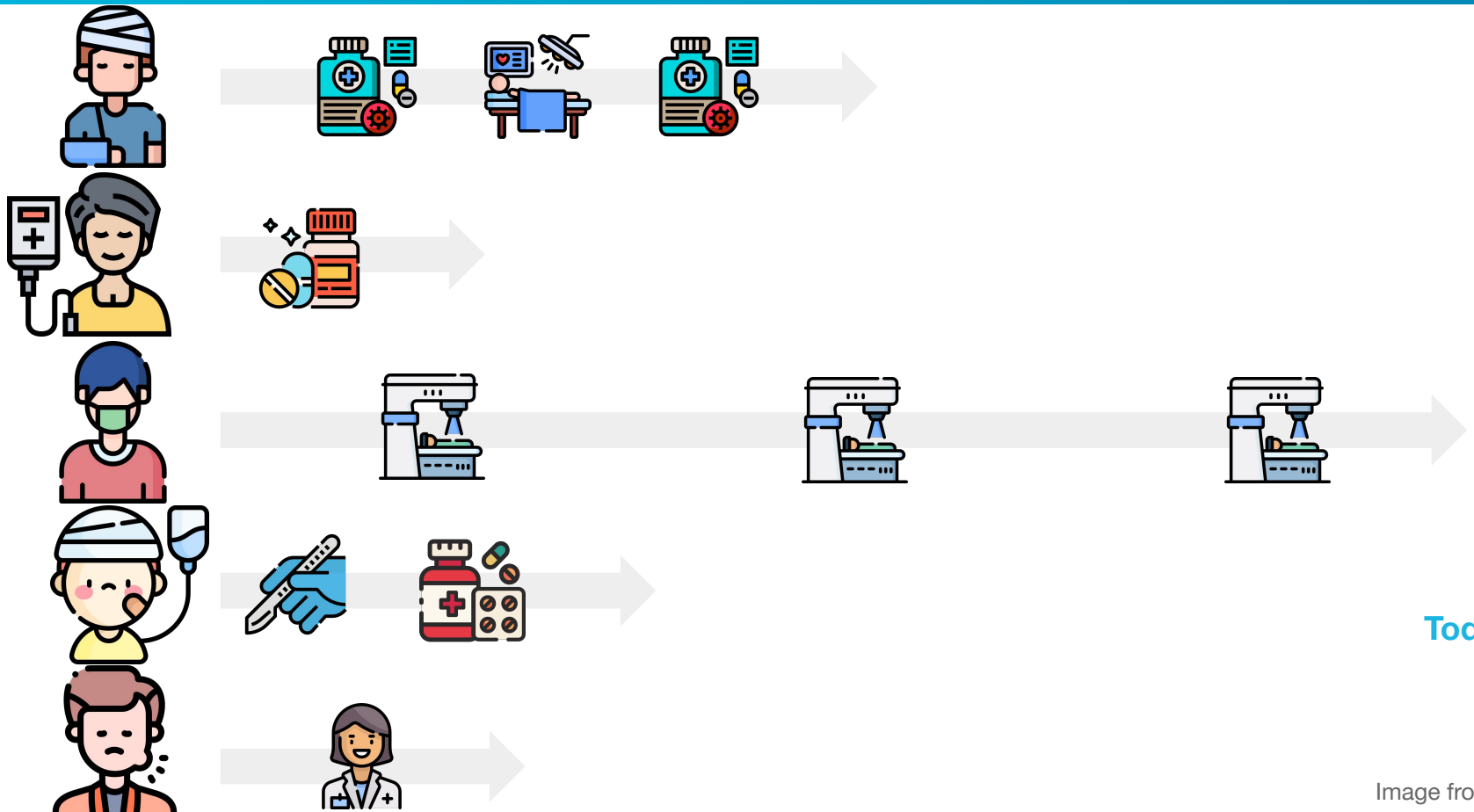
How can AI identify target patients?

How can we build a dataset for training?

Who will benefit the most?

Image from Flaticon.com

Retrospective approach



Today

Image from Flaticon.com

Dataset to study

All primary care patients enrolled in risk-based contracts from 2013 to 2015 in a large academic hospital.

- 6,079 patients who self-identified as Black (11,929 patient-years)
- 43,539 patients who self-identified as White (88,080 patient-years)

How the algorithm works

The algorithm generates “Risk scores (0 to 1)” for each patient.

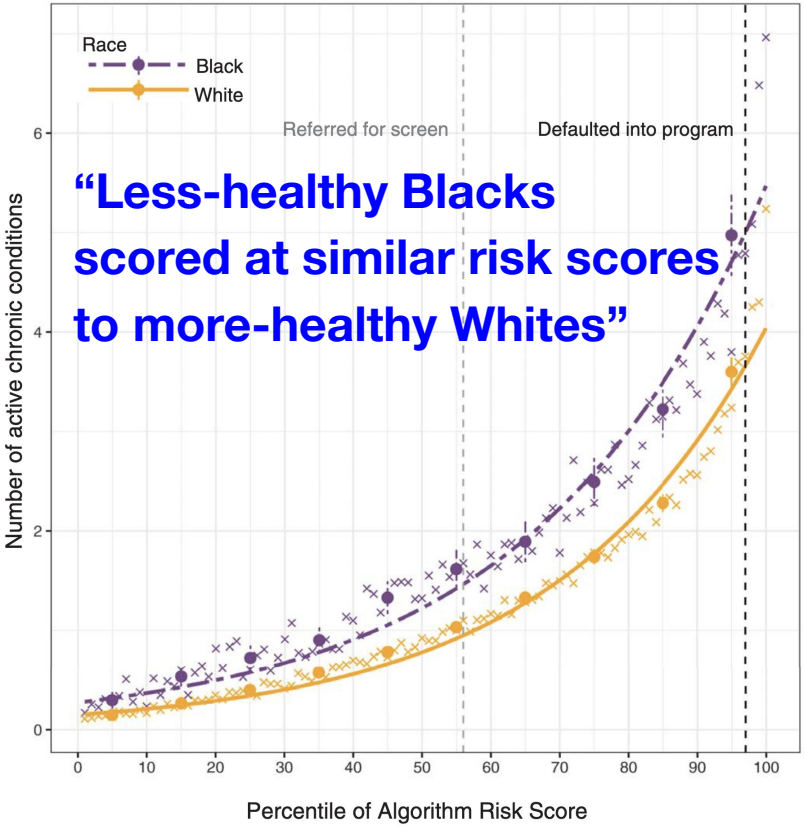
Patients ≥ 0.97 are automatically identified for enrollment in the program.

Patients ≥ 0.55 are referred to their primary care physician and asked to consider whether they would benefit from program enrollment.

Risk score vs. actual health conditions

At the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites.

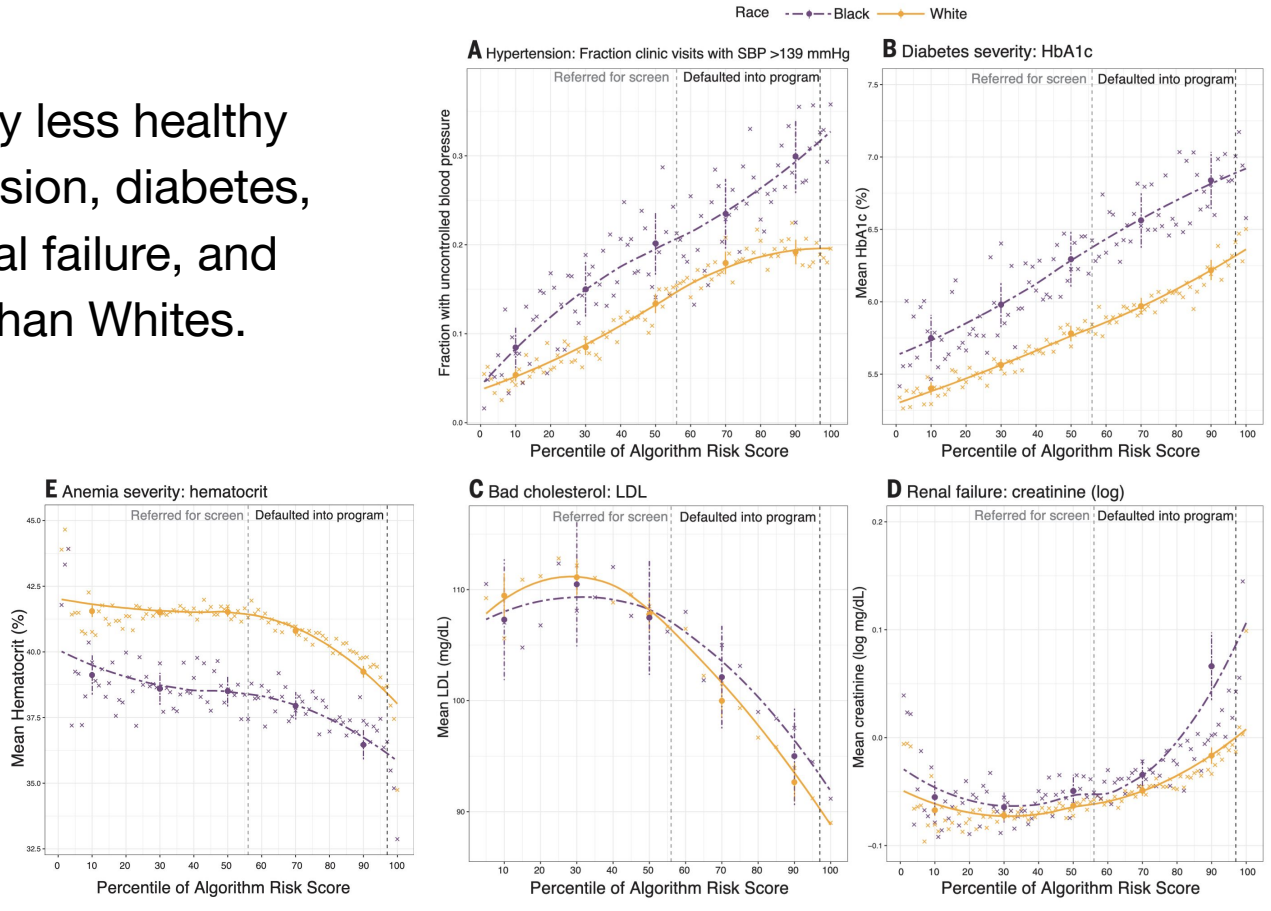
For a very-high-risk group (97th percentile), Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$)



Consistent bias across biomarkers

SMU Singapore Management University
School of Computing and Information Systems

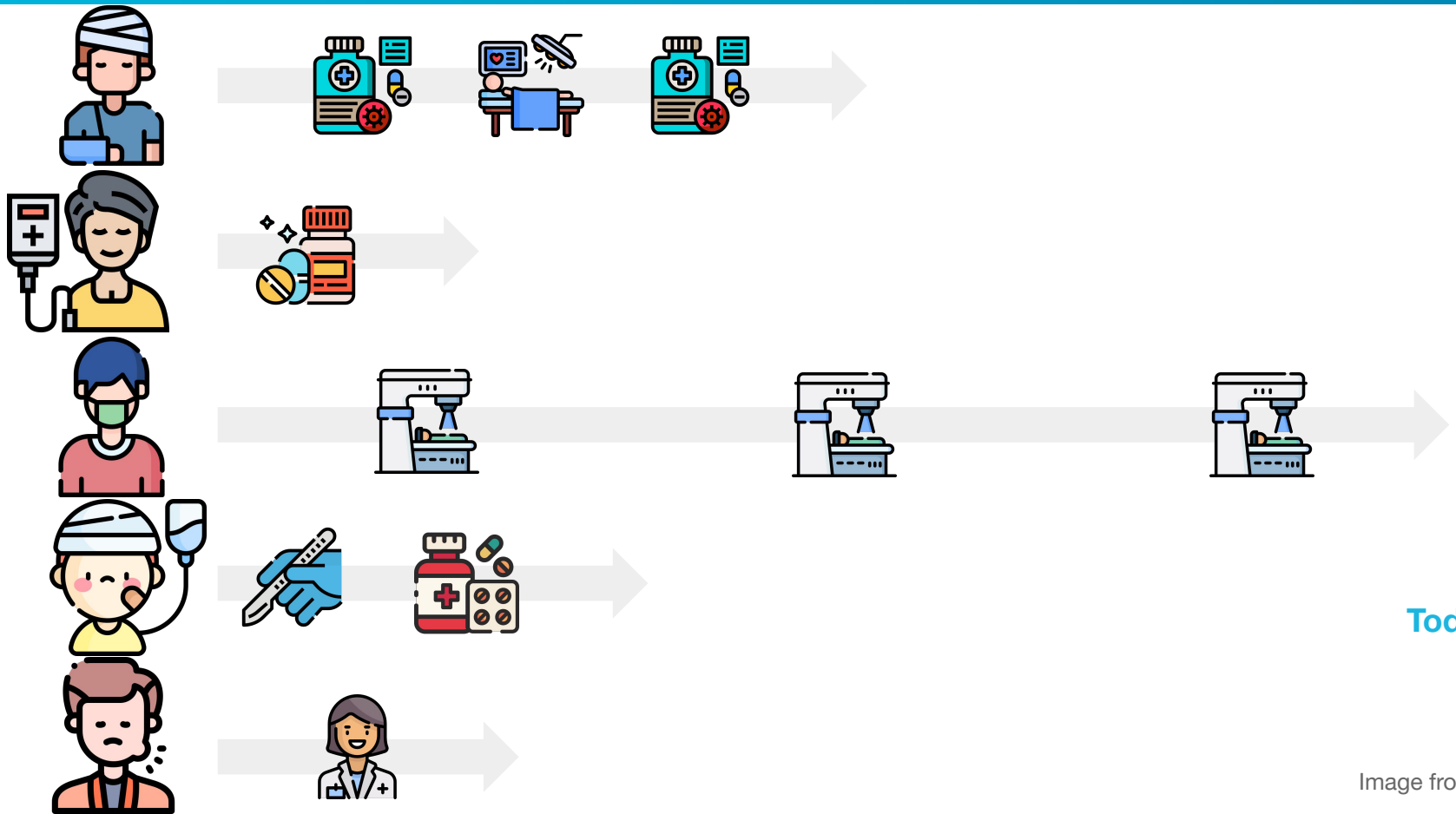
Blacks are substantially less healthy (more-severe hypertension, diabetes, higher cholesterol, renal failure, and more-severe anemia) than Whites.



Where does the racial bias come from?

Is race an input of the algorithm? No. The algorithm specifically excludes race.

How can we “compare” different treatments?



Today

Image from Flaticon.com

Expenditure of treatment for comparison

The algorithm takes total medical expenditures as the label of a patient.

- Patients with the greatest future costs could have the greatest benefit.
- The cost label reflects the industry-wide approach.
- Similar algorithms are developed and used by non-profit hospitals, academic groups, and governmental agencies (not only in industry).

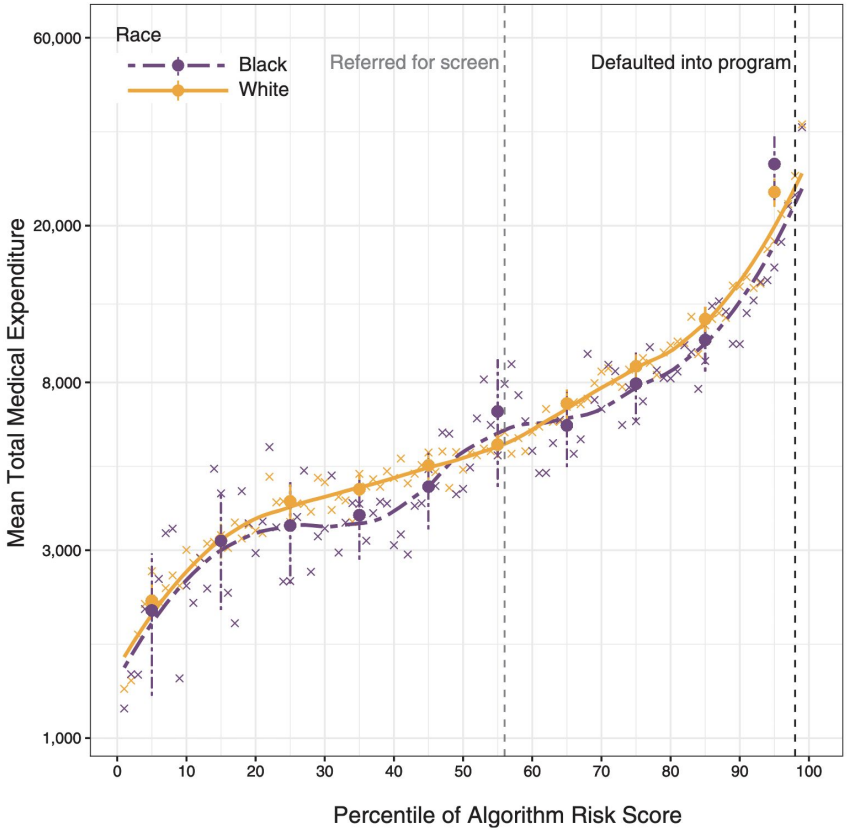
✓ The algorithm's prediction on health needs is, in fact, a prediction on how much a patient will pay.

Risk score vs. total expenditure

SMU Singapore Management University
School of Computing and Information Systems

At every level of algorithm predicted risk, Blacks and Whites have (roughly) the same costs the following year.

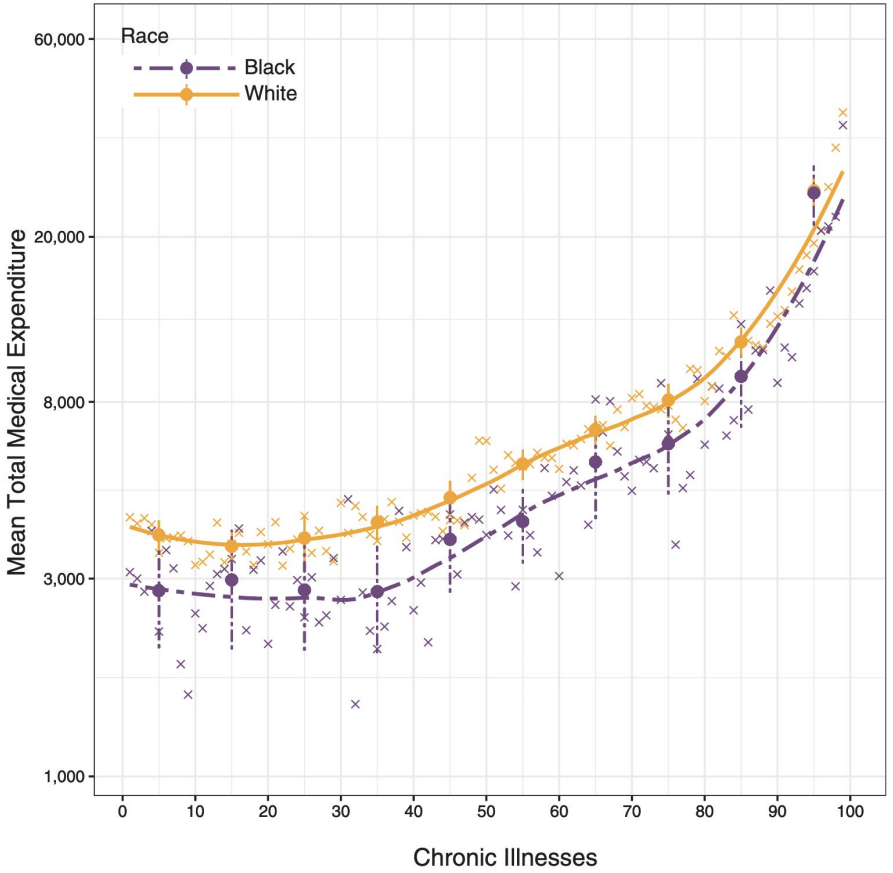
In other words, the algorithm's predictions work very well across races.



Then, why?

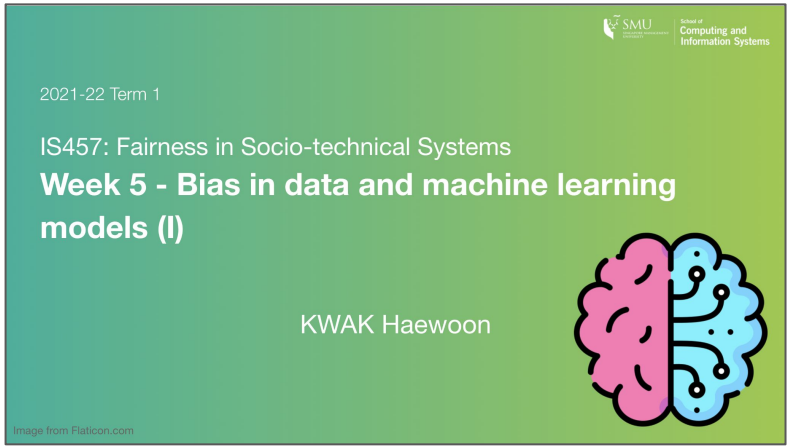
At a given level of health conditions (measured by number of chronic illnesses), Blacks show lower expenses than Whites.

Accurate prediction of costs necessarily means being racially biased on health.



Dilemma of which label to choose

“Problem formulation” in data science: the task of turning an often amorphous concept we wish to predict into a concrete variable that can be predicted in a given dataset.



After the study

The algorithm manufacturer independently replicated the same analysis on its national dataset of 3,695,943 patients and confirmed the biases.

The existing model infrastructure with the changed label (combination of cost + health condition prediction) showed an 84% reduction in bias.

Label biases are fixable, but producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment.

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]

[a]Research Institute for Signals, Systems and Computational Intelligence sinc(i), Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas CONICET, Santa Fe CP3000, Argentina; [b]Instituto de Matemática Aplicada del Litoral, Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas, Santa Fe CP3000, Argentina; and [c]Facultad de Ingeniería, Universidad Nacional de Entre Ríios, Oro Verde CP3100, Argentina

https://www.pnas.org/content/117/23/12592

AI systems for medical images

Radiographic interpretation often takes place in environments without qualified colleagues available for second opinions.

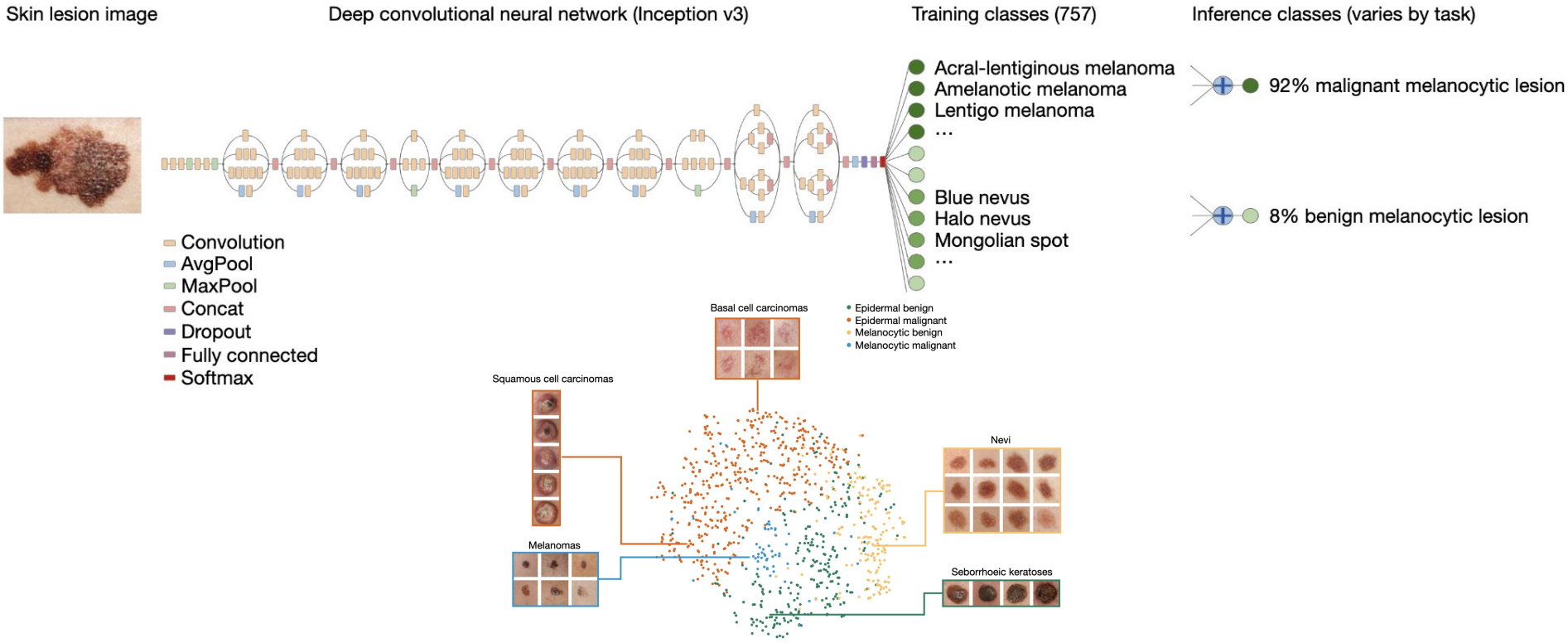
Such circumstances increase the risk of inaccurate identification of fractures on radiographs.

In emergency departments, missed fractures account for between 41 and 80% of reported diagnostic errors.

Provide useful annotations

Lindsey, Robert, et al. "Deep neural network improves fracture detection by clinicians." Proceedings of the National Academy of Sciences 115.45 (2018): 11591-11596.

Dermatologist-level classification of skin cancer



Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." nature 542.7639 (2017): 115-118.

Sex imbalance in datasets

Additionally, some diseases are more common in one sex than the other.

More males in experimental and clinical data

In early 2000, sex-specific biological differences were neglected and both experimental and clinical studies were fundamentally focused on male experimental models or male subjects.

Even nowadays, male mouse models are overall more represented than female models in basic, preclinical, and surgical biomedical research.

P21

Origin and impact of those biases

The lack of representation of female models and patients is partly due to technical and bioethical considerations

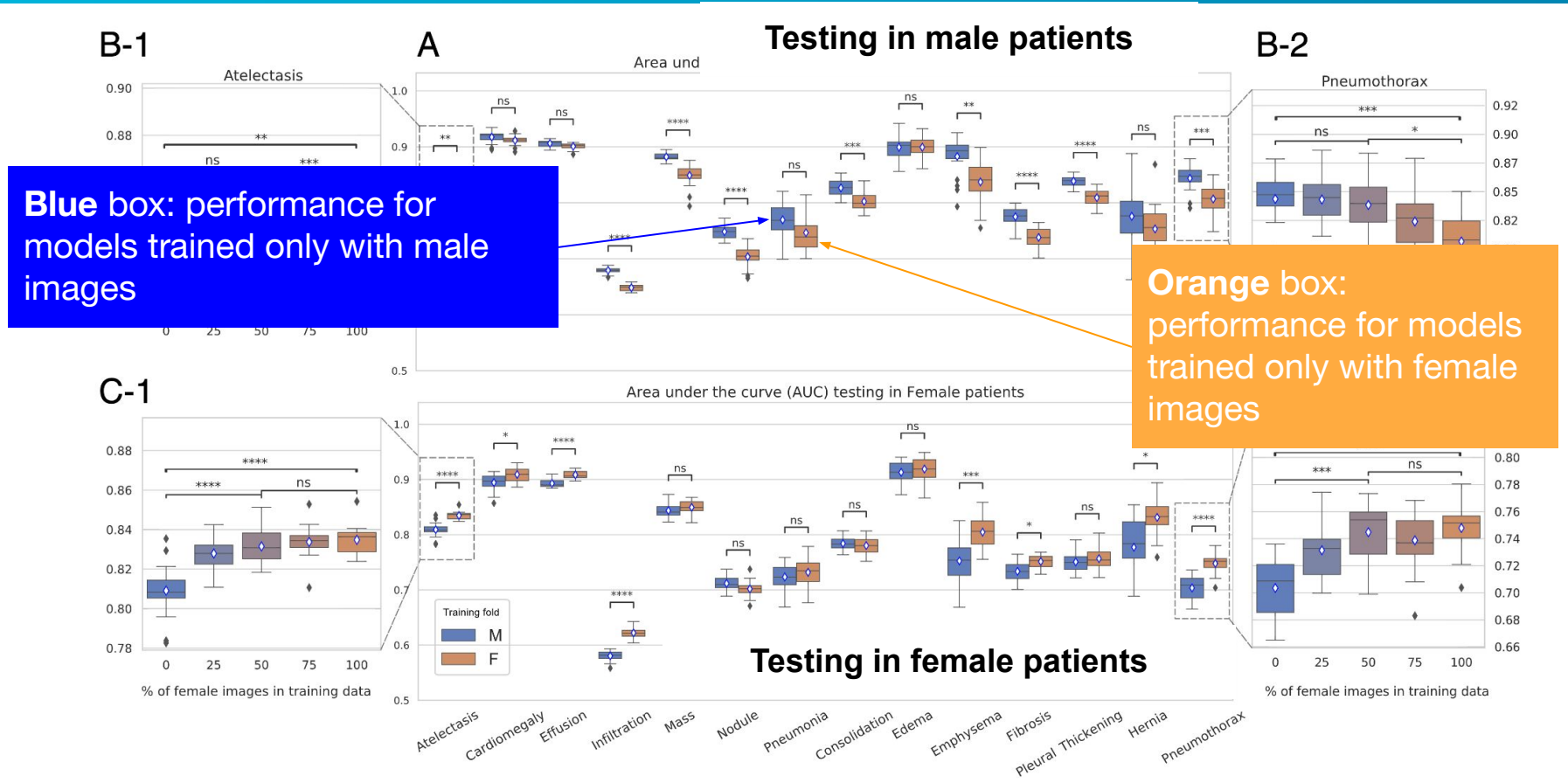
- Reduce the impact of estrous cycle in experimental studies
- Protective policies for women of childbearing age in clinical research.

Some of the treatments that currently exist for several diseases are not adequately evaluated in women who are likely to be underrepresented in clinical trials, especially in Phases I and II.

P22

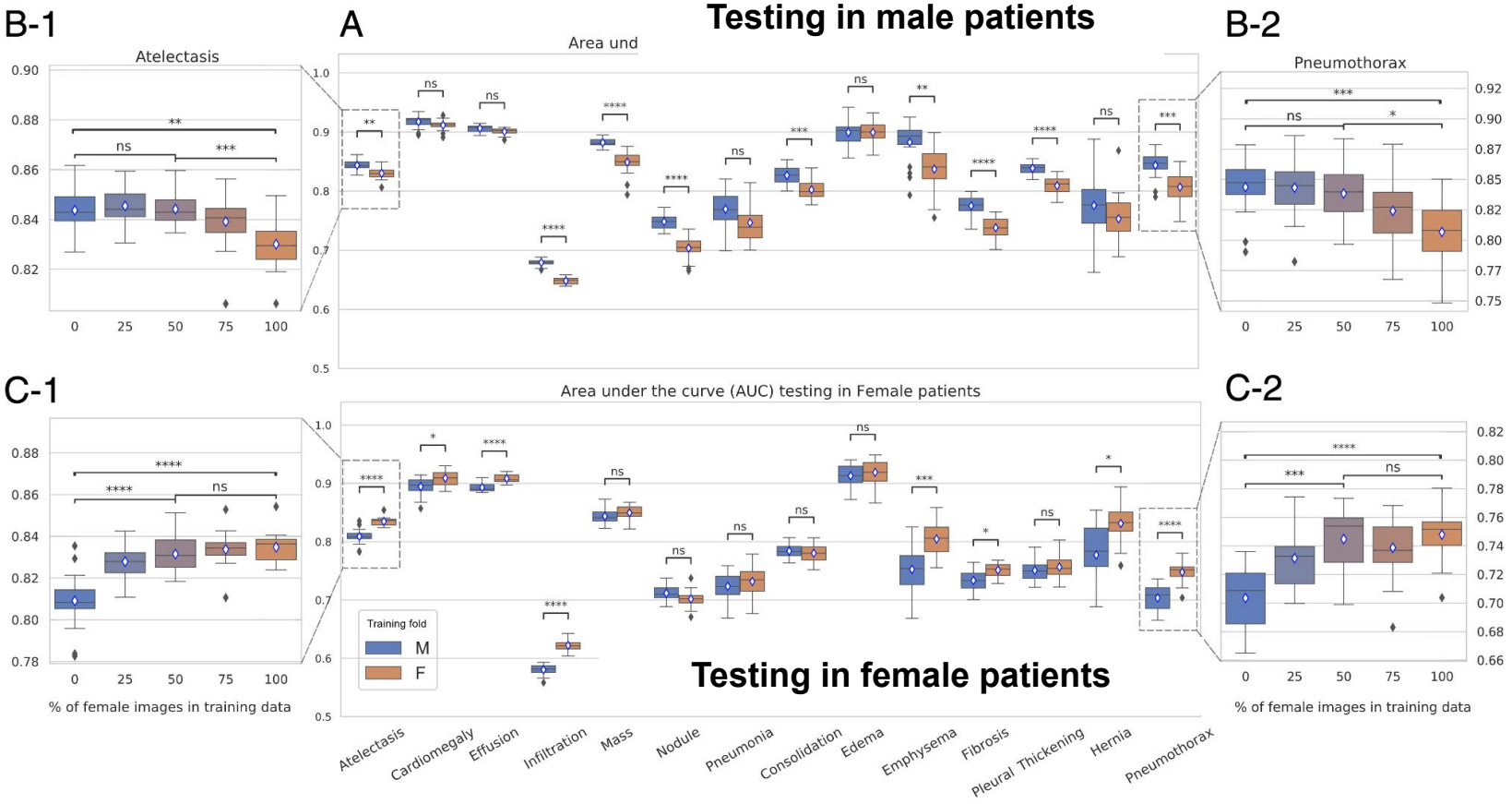
Performance with imbalanced datasets

SMU Singapore Management University
School of Computing and Information Systems



Performance with imbalanced datasets

SMU Singapore Management University
School of Computing and Information Systems



Importance of diverse/balanced dataset

Diversity should be prioritized when designing databases used to train machine learning-based computer-aided diagnosis systems.

Insufficient regulations (2020)

US Food and Drug Administration

- No explicit mention of gender/sex as one of the relevant demographic variables that should describe the sampled population.

Medical imaging community

- Most public datasets do not contain gender/sex information at the patient level (MIMIC-CXR x-ray dataset (2019), Retinal Fundus Glacucoma Challenge database (2020), etc.)

Group activity: Ethics in AI healthcare

There are no correct answers for following questions. Share your thoughts with friends.

1. Imagine a situation that a doctor uses Zoom to tell a patient she would die. After COVID-19, the broad adoption of telemedicine may make this practice common. Would you accept this practice?
2. The presence of the doctor on Zoom now changes to a “chatbot” machine. The chatbot tries to talk about end-of-life issues to patients. Would you accept this practice?
3. By analyzing big data, an AI system that can predict when a patient is going to die can be built. Do you think that using such a system is okay?
4. Where do you draw your line between “what machines (AI+robots) can do” and “what machine should not do” in healthcare domain? Why do you think so?

Ideas from https://aboutdigitalhealth.com/2021/03/30/the-ethics-of-ai-in-healthcare/

Bias in criminal justice systems

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Risk assessments become common

In every stage of the criminal justice systems, rating a defendant's risk of future crime by computer algorithms is encouraged.

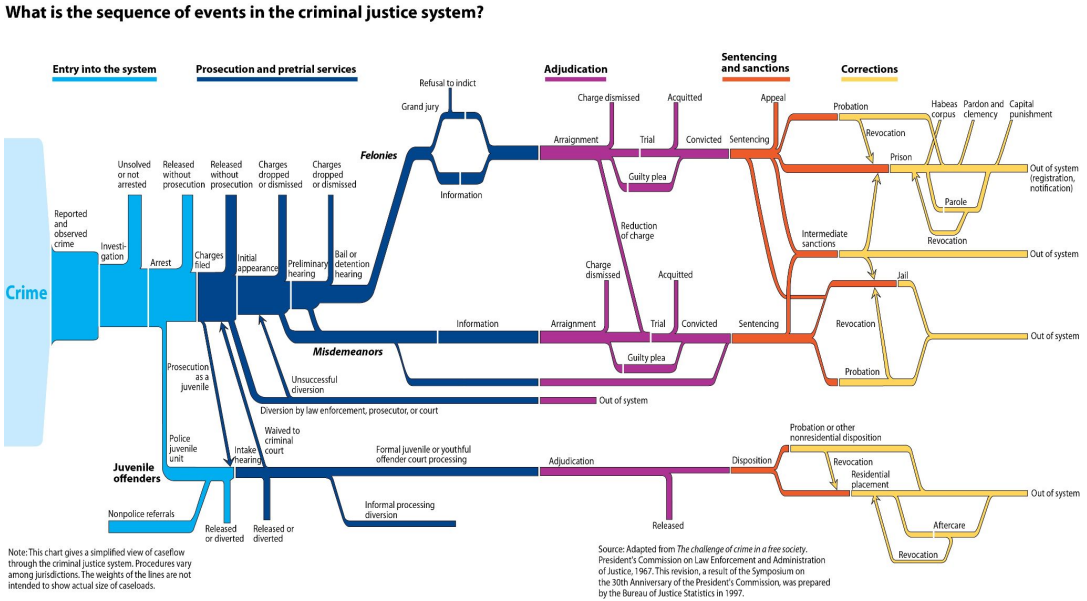
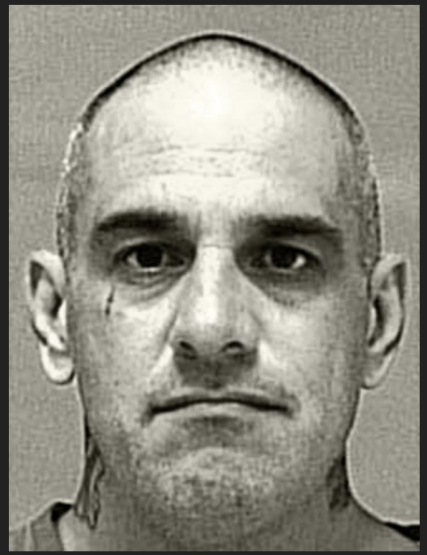


Image from https://bjs.ojp.gov/justice-system



VERNON PRATER

RISK: 3

BRISHA BORDEN

RISK: 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Analysis of risk scores by ProPublica

More than 7,000 people arrested in Broward County, Florida, in 2013 and 2014.

Checked how many were charged with new crimes over the next two years.

Risk scores are unreliable

Only 20% of the people predicted to commit violent crimes actually went on to do so.

Even when a full range of crimes were taken into account (including misdemeanors such as driving with an expired license), of those deemed likely to be reoffend, 61% were arrested for any subsequent crimes within two years.

Racial disparities in risk scores

In forecasting who would re-offend, the algorithm made mistakes with black and white defendant at roughly the same rate but:

- Falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Details of the algorithm are unknown

Northpointe's software does not publicly disclose the calculations; it is not possible to see what might be driving the racial disparity.

Only some variables are known:

- Education levels
- Whether a defendant has a job
- And more

Policing and discrimination

https://edition.cnn.com/us/live-news/george-floyd-memorial-service-protests

How predictive policing works

https://www.youtube.com/watch?v=YxvyeaL7NEM

Police records do not measure crime

Police databases are not a complete census of all criminal offences, nor do they constitute a representative random sample.

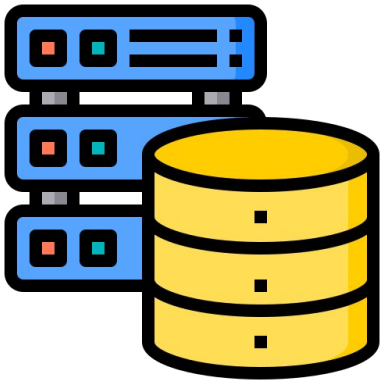
Crimes that occur in locations frequented by police are more likely to appear in the database simply because that is where the police are patrolling.

Police records capture some complex interaction between criminality and policing strategy.

Over-policing and feedback loop

Make a biased prediction based on the bias in initial training data

Patrol "predicted" areas



Police database

Police officers

Observed criminal acts

Update crime records

Strengthen beliefs

image: Flaticon.com

Over-policing and feedback loop

Officers become increasingly likely to patrol “predicted” areas and observe new criminal acts that confirm their prior beliefs regarding criminal activity.

The newly observed criminal acts then feed into the predictive policing algorithm on subsequent days, generating increasingly biased predictions.

This creates a feedback loop where the model becomes increasingly confident that the locations most likely to experience further criminal activity are exactly the locations they had previously believed to be high in crime.

Week 2 reflection

https://forms.gle/eNErPJQwKdAWfvac7