iPC Workshop on Ethical and Regulatory Issues
(2) Fairness

Michele Loi, Ph.d.

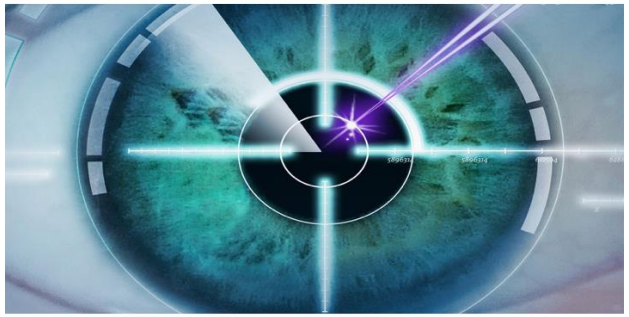
Institute of Biomedical Ethics and the History of Medicine

University of Zurich

Machine learning is used to enhance diagnosis, therapy choice, and effectiveness of the health system

Machine learning

HOME > INSIGHTS >
FDA PERMITS MARKETING OF FIRST AUTONOMOUS ARTIFICIAL INTELLIGENCE-BASED MEDICAL DEVICE



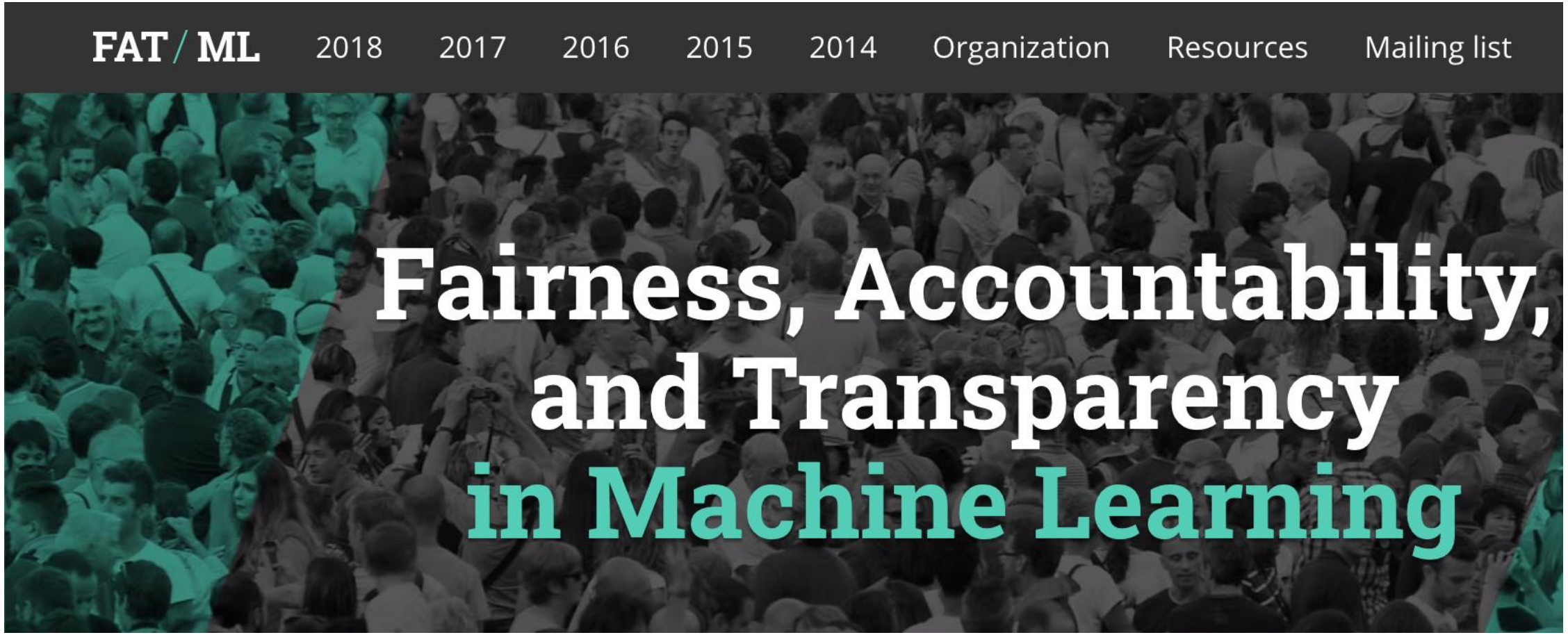
FDA Permits Marketing of First Autonomous Artificial Intelligence-Based Medical Device

MAY 2018 | ALERTS

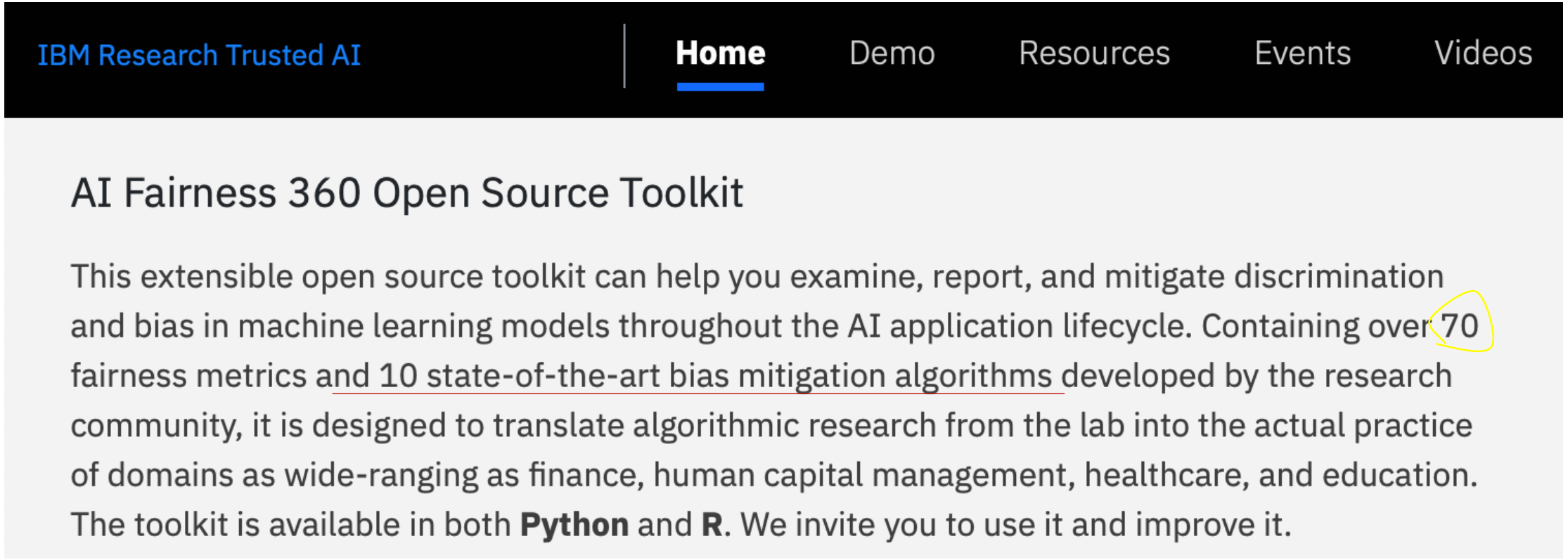
- Algorithms that can learn from large data sets to make predictions without being explicitly programmed (except in very general statistical methods - no or little domain knowledge is used)

Bias and fairness in ML

Attempts to remove bias



Many definitions of bias and “fair algorithm” exist



[e.g.](http://research.google.com/bigpicture/attacking-discrimination-in-ml/)

http://research.google.com/bigpicture/attacking-discrimination-in-ml/

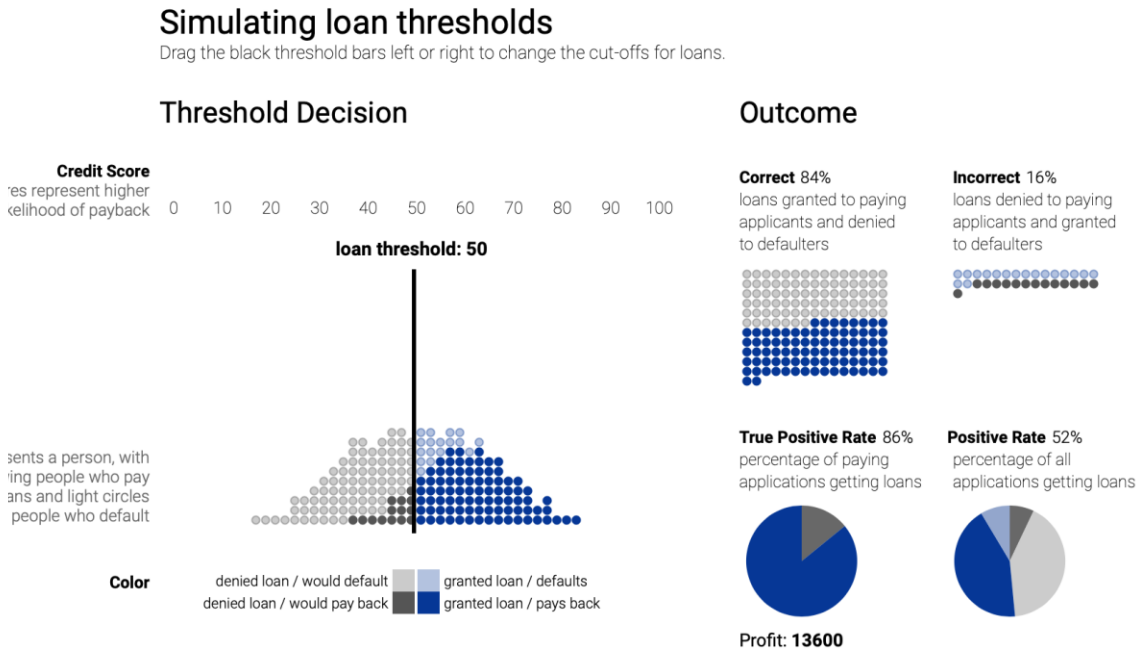
Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

This online textbook is an incomplete work in progress. Essential chapters are still missing. In the spirit of open review, we solicit broad feedback that will influence existing chapters, as well as the development of later material.

CONTENTS



Display a menu

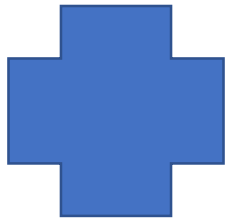
Source of bias: data?

”latent biases in training data may be perpetuated or even amplified”

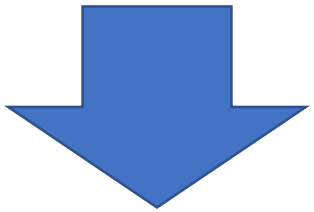
But what does it mean for data to be “biased”?

E.g.

Sub-optimal research practices (from the point of view of fairness)



Unequal access to health care (rich vs. poor, city vs. countryside, race)



Biases in the data (reflecting injustice)

Not so simple, however

Unequal baseline (e.g. colorectal cancer, more common in men than in women)

Bias?

Unequal baseline conditions

1) Are learned by ML systems capable of statistical generalization
2) Will be reflected in diagnostic systems and the decisions based on such diagnoses
3) These decisions have a tendency to appear biased, even if they are not
4) In fact, it is actually difficult to agree on whether these decisions are biased or not
5) For the same reason, it is difficult to disagree on whether *removing* the alleged bias makes the decisions fairer (or less fair) instead
6) This is an instance of reasonable disagreement – different people may reasonably disagree. It is ultimately a *moral* not a technical question whether something is a bias or not.

Case study

A hospital has 840 male patients and 840 female patients.

Of these, 252 males and 189 females have cancer.

These data are used to train a machine learning system.

(Notice, colorectal cancer tends to be more common in men than women, generally, in the population)

The machine learning system is used to make an early, low accuracy, prediction, which is used to prioritize access to very accurate, but also very expensive, clinical tests.

uzh.voting/hetw

A question to the audience

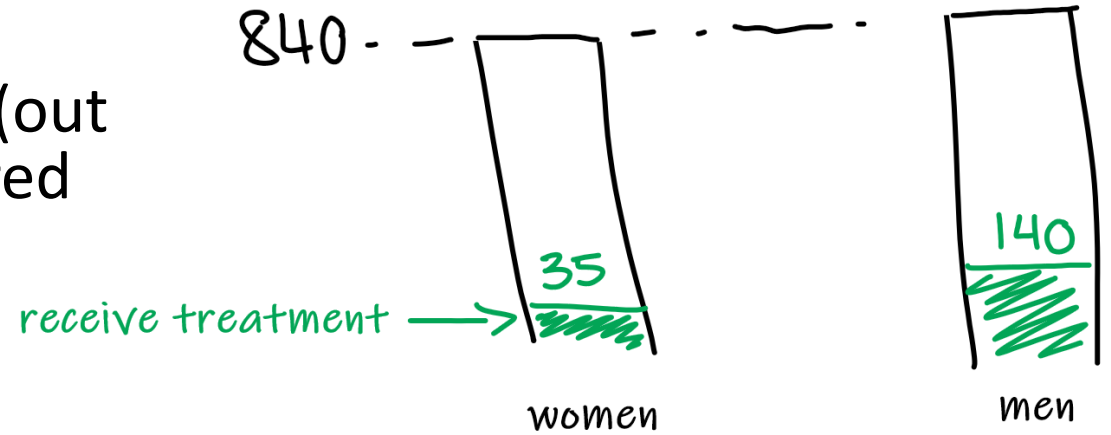
- Participate in the **poll**, if you can. Otherwise answer the question in your head.

Suppose that the *predicted cases of possible colorectal cancer are as follows:*

- *Women: 35*
- *Men: 140*

It means that only 35 women (out of 840 women patients), compared to 140 men (out of the same number of men) will be offered expensive clinical testing.

Is this necessarily unfair?



Possible reply: men, on average, need testing more than women do, so it is not necessarily unfair if a higher proportion of men gets access to the expensive/accurate clinical testing

- Problem: what is need, statistically speaking?
- There are, *at least*, two equally plausible *prima facie,* statistical interpretations of what this “equal need” is.
- These interpretations can be *coded* in the algorithm to make it fair, e.g. through the algorithms mentioned above.
- But they cannot be achieved simultaneously (only very poorly approximated)

Fairness 1: predictive value parity

The probability of a true/false prediction is statistically independent from sex

	Same in model I and II	True labels (real cases of cancer)				Model I	True labels (real cases of cancer)		
		Cancer	No cancer				Cancer	No cancer	
Predicted labels (predicted cases of cancer)	Cancer	**(a)** 112	**(b)** 28	140	Predicted labels (predicted cases of cancer)	Cancer	**(a)** 28	**(b)** 7	35
	No cancer	**(c)** 140	**(d)** 560	700		No cancer	**(c)** 161	**(d)** 644	805
	Total patients	252	588	840		Total patients	189	651	840

Table 1. Model I equalizes the ratios $\frac{a}{a+b}$ *and* $\frac{d}{c+d}$ *for men and women (predictive value parity).*

Positive predictive value (men) = $\frac{112}{112+28} = 0.8$ *Negative predictive value (men) =* $\frac{560}{140+560} = 0.8$

Positive predictive value (women) = $\frac{28}{28+7} = 0.8$ *Negative predictive value (women) =* $\frac{644}{161+644} = 0.8$

Moral interpretation

If you think that groups with the same average risk have equal needs

And

people with equal needs should have the same probability of obtaining the expensive/accurate clinical examination, *then*

Since:

- *Man predicted to be positive** AND
- *Woman predicted to be positive**

(a) Have the same average risk (0.8)

(b) Obtain access the expensive/accurate examination

This is fair.

ditto* for **negative

Fairness 2: equal true/false positive and true/false negative rate

The probability for a true positive/negative to be correctly identified is statistically independent from sex

	**Same in model I and II**	**True labels (real cases of cancer)**				**Model II**	**True labels (real cases of cancer)**		
		Cancer	No cancer				Cancer	No cancer	
Predicted labels (predicted cases of cancer)	Cancer	**(a)** 112	**(b)** 28	140	Predicted labels (predicted cases of cancer)	Cancer	**(a)** 84	**(b)** 31	115
	No cancer	**(c)** 140	**(d)** 560	700		No cancer	**(c)** 105	**(d)** 620	725
	Total patients	252	588	840		Total patients	189	651	840

Table 2. Model II equalizes the ratios $\frac{a}{a+c}$ *and* $\frac{d}{b+d}$ *for men and women (equalized odds).*

True positive rate (men) = $\frac{112}{112+140} \approx 0.44$ *True negative rate (men) =* $\frac{560}{28+560} \approx 0.95$

True positive rate (women) = $\frac{84}{84+105} \approx 0.44$ *True negative rate (women) =* $\frac{620}{31+620} \approx 0.95$

Moral interpretation

If you think that groups with the same ~~average risk~~ disease have equal needs

And

people with equal needs should have the same probability of obtaining the expensive/accurate clinical examination, *then*

Since:

- *True positive* men* AND
- *True positive* women*

    (a) Have the disease

    (b) Are equally likely to obtain access the expensive/accurate examination

So, this is fair.

ditto* for **negative

Can predictive value parity and equalized odds (true/false positive/negative rates) be *both fair*

Mathematically = this is often impossible (with rare exceptions)

See: Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv:1610.07524 [Cs, Stat]*. http://arxiv.org/abs/1610.07524

Philosophically = this corresponds to two *alternative* interpretations of who needs the examination the most:

> need (1): people equally at risk are those who should have equal chances of obtaining the examination
>
> need (2): people with the disease are those who should have equal chances of obtaining the examination

The virtual patient case study

A hospital has 840 white children and 840 black children.

The machine learning system is used to make a prediction based on a statistical model of the patient built on his the omics data (virtual patient), which is used to decide if the therapy works on the child.

Suppose that the therapy is more successful - on average - for white children than black children.

A similar problem emerges…

1) Is it fair for black children to be less likely to obtain the therapy, when the baseline for success of the therapy is different in the two populations?

Notice

- This evaluation of what is fair requires a careful evaluation of the benefits and harms, e.g.

- A) what are the side-effects of the therapy?
- B) is it more harmful to be a false negative (to not be given a therapy that could work) or to be a false positive (to be given a therapy that does not work and may have side effects)?

The same dilemma

- 2. The choice between equalizing the true(false) positive rate vs. equalizing predictive value is not ethically trivial and does not have (yet) a clear ethical answer

Fairness 1: predictive value parity

The probability of a true/false prediction is statistically independent from being white / black

White children	Same in model I and II	True labels (realityr)			Black children	Model I	True labels (reality)		
		Therapy works	Therapy does not work				Therapy works	Therapy does not work	
Predicted labels (expectations)	Therapy works	(a) 112	(b) 28	140	Predicted labels (expectations)	Therapy works	(a) 28	(b) 7	35
	Therapy does not work	(c) 140	(d) 560	700		Therapy does not work	(c) 161	(d) 644	805
	Total patients	252	588	840		Total patients	189	651	840

Table 1. Model I equalizes the ratios $\frac{a}{a+b}$ *and* $\frac{d}{c+d}$ *for white and black children (predictive value parity).*

Positive predictive value (white) = $\frac{112}{112+28} = 0.8$ *Negative predictive value (white) =* $\frac{560}{140+560} = 0.8$

Positive predictive value (black) = $\frac{28}{28+7} = 0.8$ *Negative predictive value (black) =* $\frac{644}{161+644} = 0.8$

Fairness 2: equal true/false positive and true/false negative rate

The probability for a true positive/negative to be correctly identified is statistically independent from being white /black

White children	Same in model I and II	True labels (real cases of cancer)			Black children	Model II	True labels (real cases of cancer)		
		Therapy works	Therapy does not work				Therapy works	Therapy does not work	
Predicted labels (expectations)	Therapy works	(a) 112	(b) 28	140	Predicted labels (expectations)	Therapy works	(a) 84	(b) 31	115
	Therapy does not work	(c) 140	(d) 560	700		Therapy does not work	(c) 105	(d) 620	725
	Total patients	252	588	840		Total patients	189	651	840

Table 2. Model II equalizes the ratios $\frac{a}{a+c}$ *and* $\frac{d}{b+d}$ *for white and black children (equalized odds).*

True positive rate (white) = $\frac{112}{112+140} \approx 0.44$ *True negative rate (black) =* $\frac{560}{28+560} \approx 0.95$

True positive rate (white) = $\frac{84}{84+105} \approx 0.44$ *True negative rate (black) =* $\frac{620}{31+620} \approx 0.95$

In the literature

- Very little moral guidance on the choice between equal odds and predictive parity
- Fairness as fair distribution of utility from the decision
  - Heidari, H., Ferrari, C., Gummadi, K., & Krause, A. (2018). Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 1265–1276). Curran Associates, Inc. http://papers.nips.cc/paper/7402-fairness-behind-a-veil-of-ignorance-a-welfare-analysis-for-automated-decision-making.pdf
  - Heidari, H., Loi, M., Gummadi, K. P., & Krause, A. (2019). A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 181–190. https://doi.org/10.1145/3287560.3287584
  - For health care: Pfohl, S., Duan, T., Ding, D. Y., & Shah, N. H. (2019). Counterfactual Reasoning for Fair Clinical Risk Prediction. *ArXiv:1907.06260 [Cs, Stat]*. http://arxiv.org/abs/1907.06260
- Several other approaches and definitions of fairness, e.g. individual fairness, counterfactual fairness, etc, not all applicable to all contexts

Ethical issues in machine learning bias in clinical prediction (general)

- David W. Bates et al., “Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients,” *Health Affairs (Project Hope)* 33, no. 7 (July 2014): 1123–31, https://doi.org/10.1377/hlthaff.2014.0041;
- I. Glenn Cohen et al., “The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care,” *Health Affairs* 33, no. 7 (July 1, 2014): 1139–47, https://doi.org/10.1377/hlthaff.2014.0048.
- Benjamin A. Goldstein et al., “Opportunities and Challenges in Developing Risk Prediction Models with Electronic Health Records Data: A Systematic Review,” *Journal of the American Medical Informatics Association: JAMIA* 24, no. 1 (2017): 198–208, https://doi.org/10.1093/jamia/ocw042;
- Alvin Rajkomar et al., “Scalable and Accurate Deep Learning with Electronic Health Records,” *Npj Digital Medicine* 1, no. 1 (May 8, 2018): 18, https://doi.org/10.1038/s41746-018-0029-1.
- Alvin Rajkomar et al., “Ensuring Fairness in Machine Learning to Advance Health Equity,” *Annals of Internal Medicine* 169, no. 12 (December 18, 2018): 866, https://doi.org/10.7326/M18-1990;
- Danton S. Char, Nigam H. Shah, and David Magnus, “Implementing Machine Learning in Health Care — Addressing Ethical Challenges,” *New England Journal of Medicine* 378, no. 11 (March 15, 2018): 981–83, https://doi.org/10.1056/NEJMp1714229;
- Milena A. Gianfrancesco et al., “Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data,” *JAMA Internal Medicine* 178, no. 11 (November 1, 2018): 1544–47, https://doi.org/10.1001/jamainternmed.2018.3763;
- Tiffany C. Veinot, Hannah Mitchell, and Jessica S. Ancker, “Good Intentions Are Not Enough: How Informatics Interventions Can Worsen Inequality,” *Journal of the American Medical Informatics Association: JAMIA* 25, no. 8 (August 1, 2018): 1080–88, https://doi.org/10.1093/jamia/ocy052;
- Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen, “Machine Learning in Medicine: Addressing Ethical Challenges,” *PLOS Medicine* 15, no. 11 (November 6, 2018): e1002689, https://doi.org/10.1371/journal.pmed.1002689.

Fairness dilemma and trade offs

Basic readings on fairness trade-offs:

- Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. *ArXiv:1703.09207 [Stat]*. http://arxiv.org/abs/1703.09207
- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *ArXiv:1609.05807 [Cs, Stat]*, September 19, 2016, http://arxiv.org/abs/1609.05807.