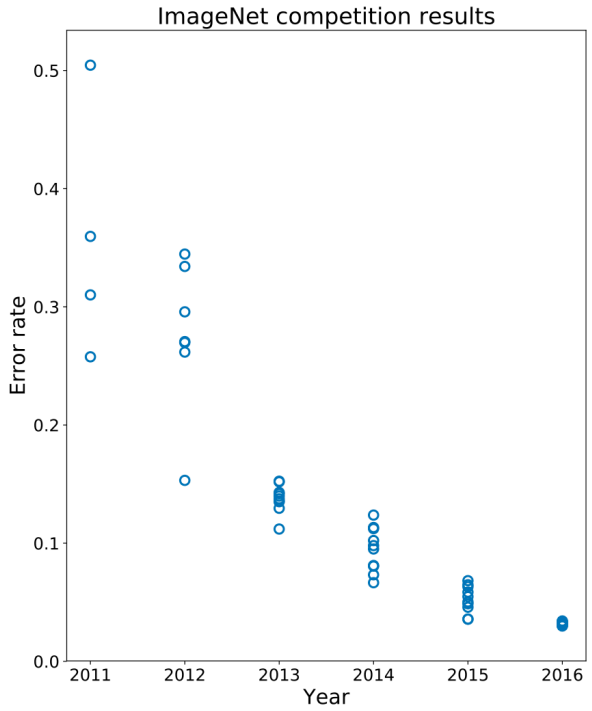
The need for clinical (and trialist) commonsense in AI algorithm design

Samuel Finlayson

MD-PhD Candidate, Harvard-MIT

We’re all really excited about machine learning, and we should be.

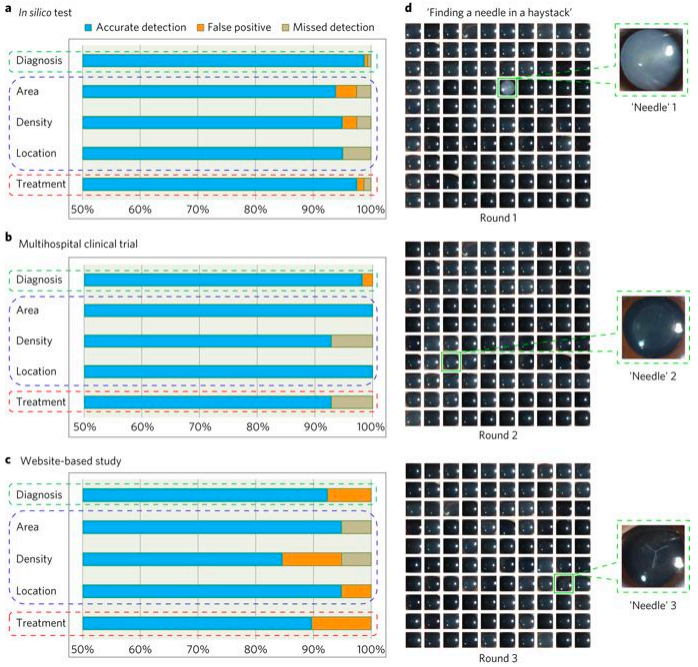


en.wikipedia.org/wiki/File:ImageNet_error_rate_history_(just_systems).svg

Source: eyediagnosis.net

For the all excitement, clinical benefit of AI is still largely hypothetical

- Very few **prospective** trials of medical AI have been reported in any specialty
  - Per Eric Topol Review, only 4 as of 1/2019
  - Good news: 2 of 4 were in Ophthalmology!
- Many models struggle to reproduce findings in **new patient populations**
- No trials, to my knowledge, have demonstrated improved clinical ***outcomes***

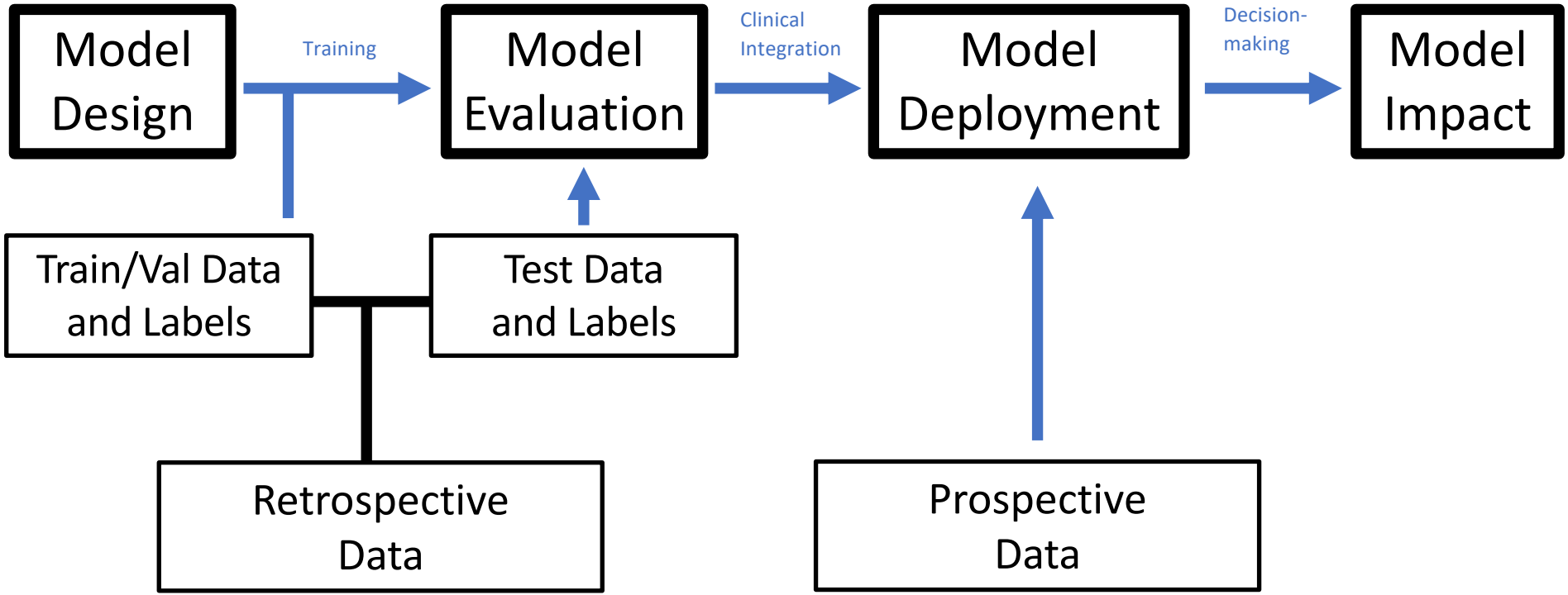


CC-Cruiser: 98.87% accuracy in small trial 1
87.4% (vs physician 99.1%) in trial 2

Goal for this tutorial:

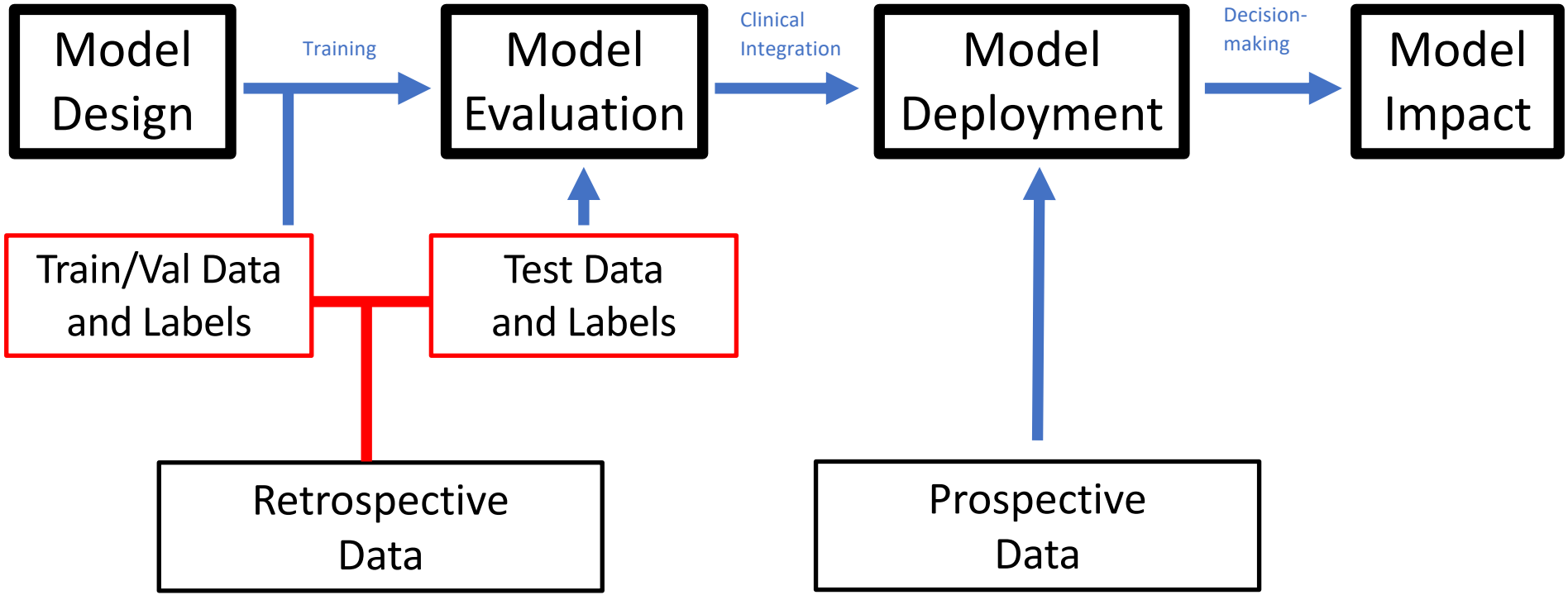
Equip attendees to identify **common pitfalls** in medical AI that make **informed clinical experts** essential to development and deployment

Review: The ML development pipeline



What do we need clinical experts to be asking?

Key questions to ask during **dataset curation**



How might our model be tainted with **information from the future**?

<u>Hypothetical example #1:</u>

- Plan: train a ML algorithm to detect DR
- Postdoc downloads all fundus images from your clinical database, using **discharge diagnoses** to gather DR cases and healthy controls.

What could go wrong?
(Hint: see figure)

Source: endotext.com

How might our model be tainted with **information from the future**?

Answer:

- *Laser scars* are present!
- Model may learn to “diagnose” the *treatment* instead of the *disease.*
- This is one example of **label leakage**, a very common problem.

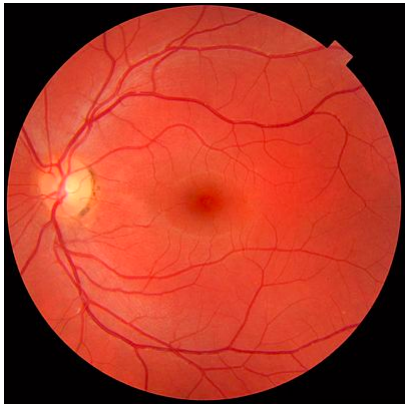
Source: endotext.com

How might our **test set** be contaminated with information from our **training set**?

<u>Hypothetical example #1 (con't):</u>

- Postdoc tries again, limiting images to exams prior to treatment.
- All case and control images split randomly into a train and test set

Training Image 1



Test Image 1

What could go wrong?
(Hint: see figure)

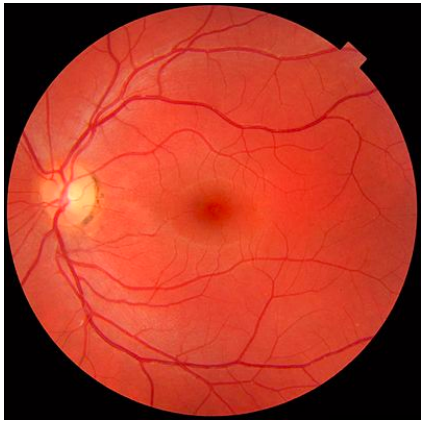
Image source: wikipedia

How might our **test set** be contaminated with information from our **training set**?

Answer:

- Images from the same patients are in both train and test sets!
- Test set metrics will *overestimate* model accuracy, providing limited evidence for accuracy on *unseen patients*
- This is one example of **train-test set leakage.**

Training Image 1



Test Image 1

Image source: wikipedia

How might our model by **confounded**?

Hypothetical example (#2):

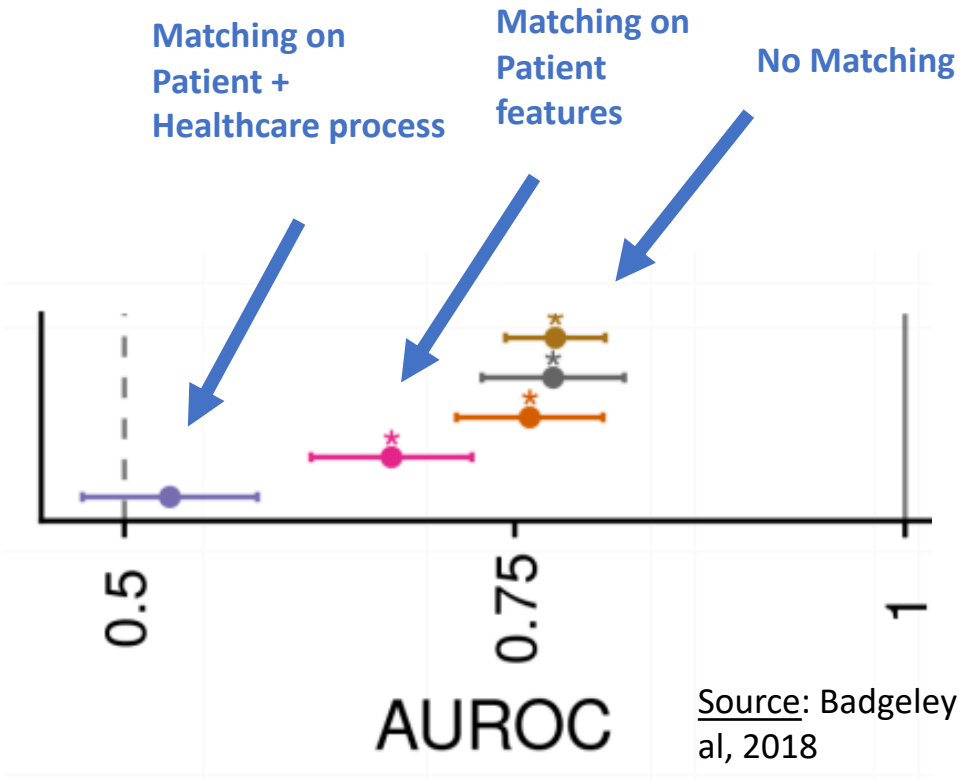
- You build an ML classifier to detect optic disk edema for neurologic screening.
- Images are gathered from the **ED** and the **outpatient clinic** with no regard to their site of origin.

How could this data acquisition process lead to **confounding**?

How might our model by **confounded**?

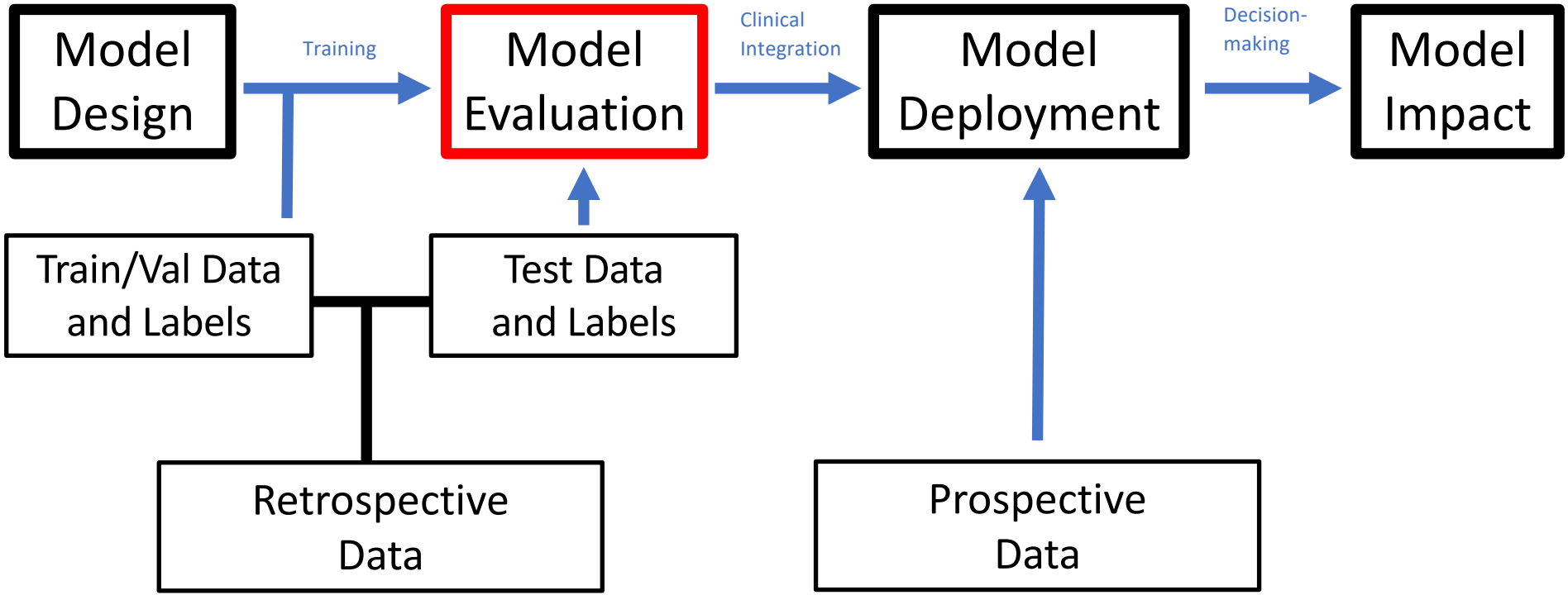
(One) Answer:

- Imaging models have been shown to depend on **"non-imaging"** variables
- In **ophthalmology**, we know that age, sex, etc. trivially predicted by models from images.
- Problem very acute with drug, billing, text data



Source: Badgeley et al, 2018

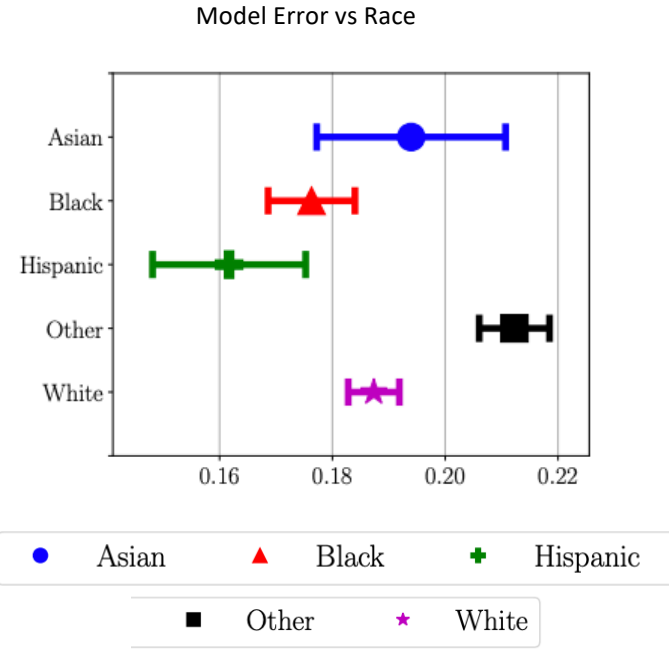
Key questions to ask during **model evaluation**



Is our model performance consistent across patient subpopulations?

Hypothetical example #3:

- At the request of reviewer #2, your team evaluates its model performance stratified by race, finding large differences. (See plot on right).
- You gather more cases from underrepresented groups and retrain the model, but it doesn't improve the situation.



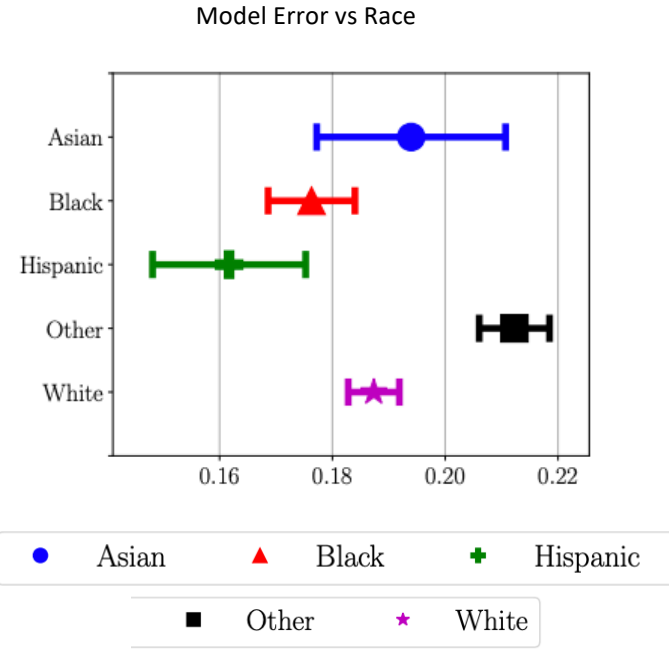
What could be happening?

Source: Chen et al, NeurIPS '18

Is our model performance consistent across subpopulations?

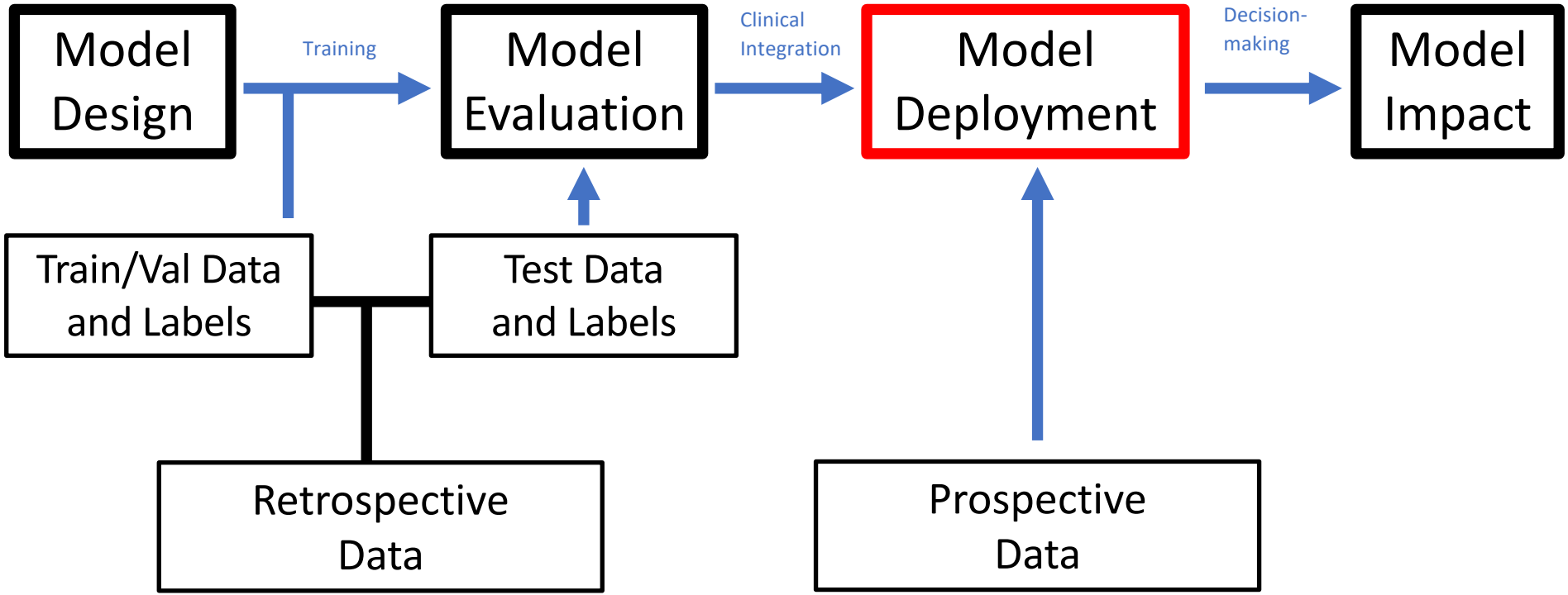
Answer:

- All model bias is not created equal
- Different biases require **different solutions**
- Could require: More **data**, more **features**, or different **models**.
- See the brilliant Chen et al, NeurIPS 2018



Source: Chen et al, NeurIPS '18

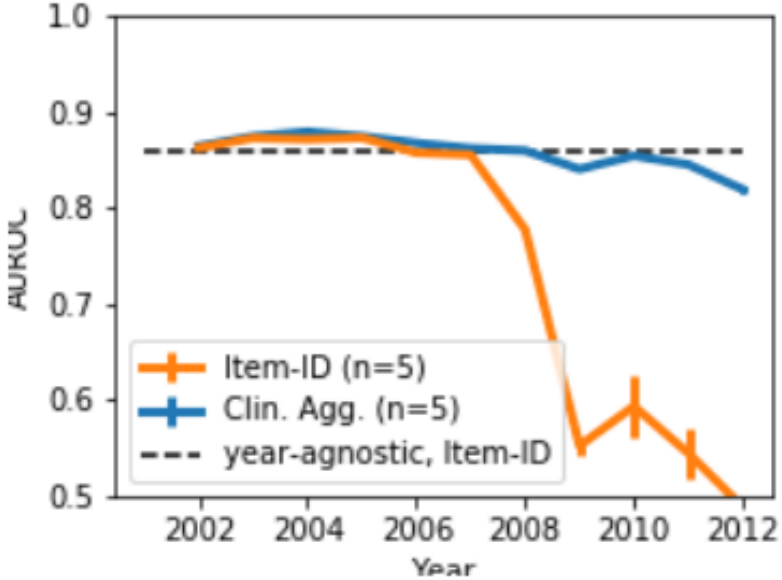
Key questions to ask during **model deployment**



How might the data we feed our model change **over time**?

<u>Hypothetical example #4:</u>

- Your highly accurate ML tool suddenly begins to fail several years after clinical deployment
- IT team insists the model has not changed.

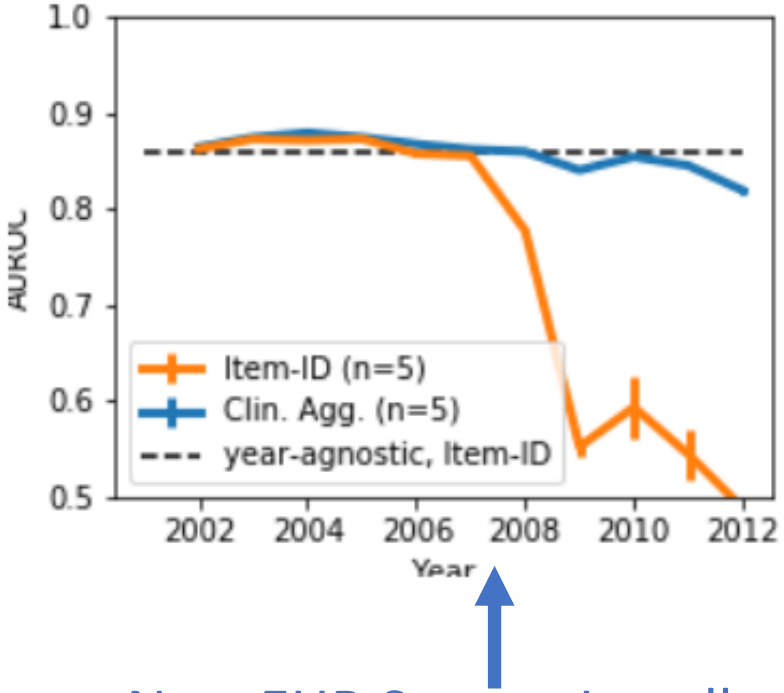


What might be going on?

How might the data we feed our model change **over time**?

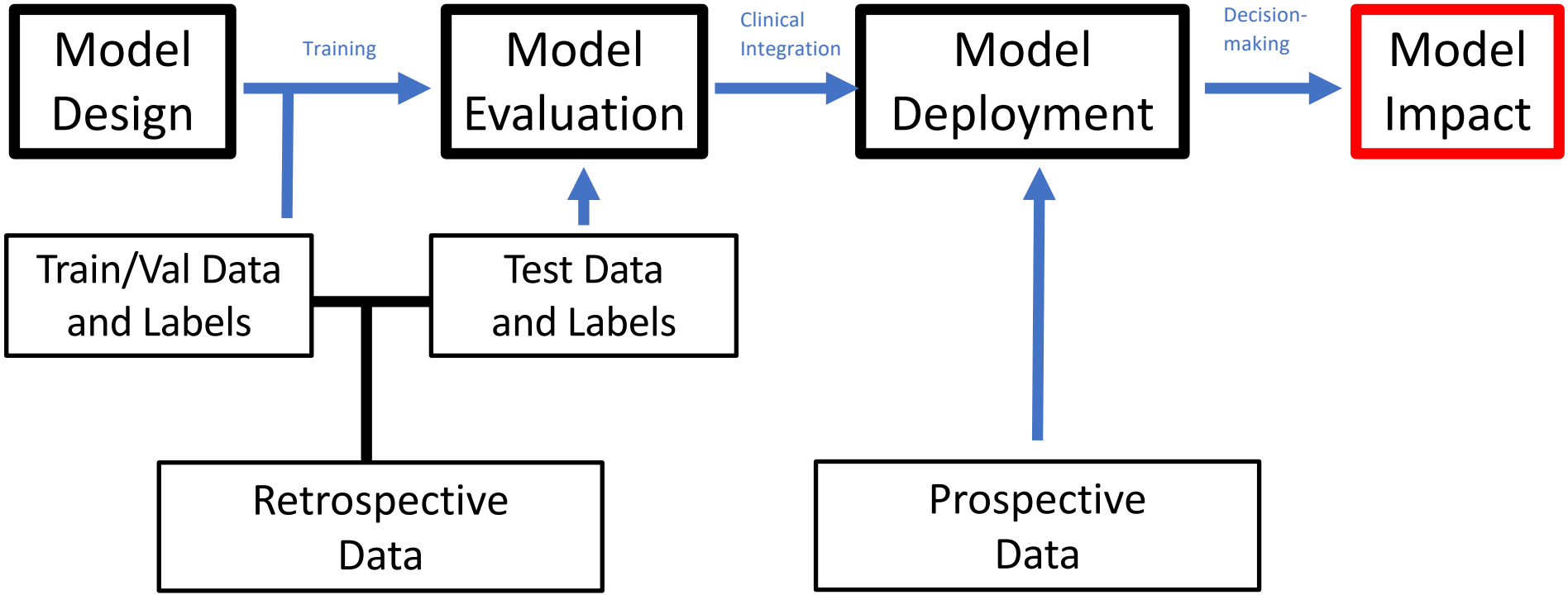
Answer:

- Clinical performance is **not fixed!**
- Changes in the input data can disrupt model performance: **dataset shift**
- Model evaluation and development must be an **ongoing**



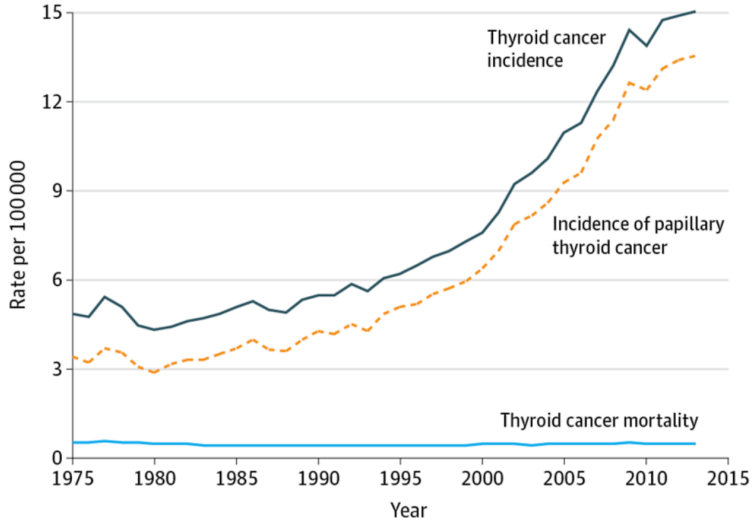
Source: Nestor et al, 2018

Key questions to ask as we assess **model impact**



Can we anticipate any unintended consequences?

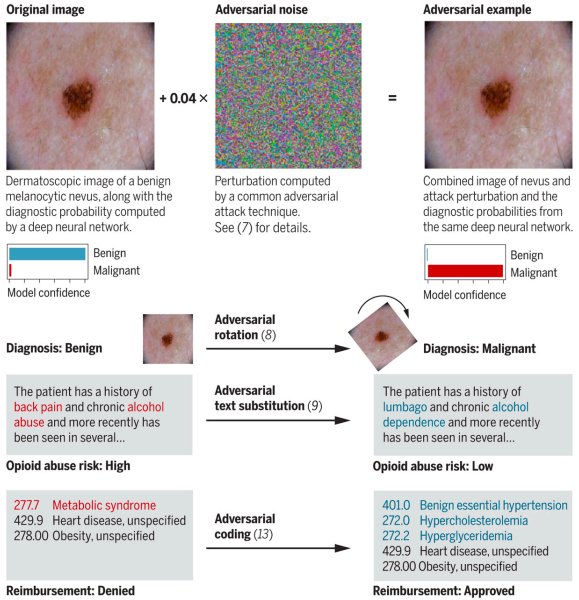
Diagnosis does not equal outcomes!



Thyroid Cancer Incidence and Mortality in the United States

Welch, 2017

Mismatched incentives -> adversarial behavior



Finlayson et al, 2019

Conclusions

- Many of the most pernicious challenges of medical machine learning are *study design problems*
  - What sources of **leakage**, **bias** and **confounding** might be baked into the design?
  - How does the **target** population compare with the **study** population?
  - How might populations **evolve over time**, and how should they be **monitored**?
  - Can we anticipate any **unintended consequences** of deployment?
- Clinicians and clinical researchers (trialists, epidemiologists, biostatisticians) have been asking similar questions for *decades*
- Delivering on the promise of medical ML requires a true partnership between clinical research and machine learning expertise

Thank you

Invitation to speak: Michael Abramoff

Feedback on presentation: Lab team of Isaac Kohane, DBMI at Harvard