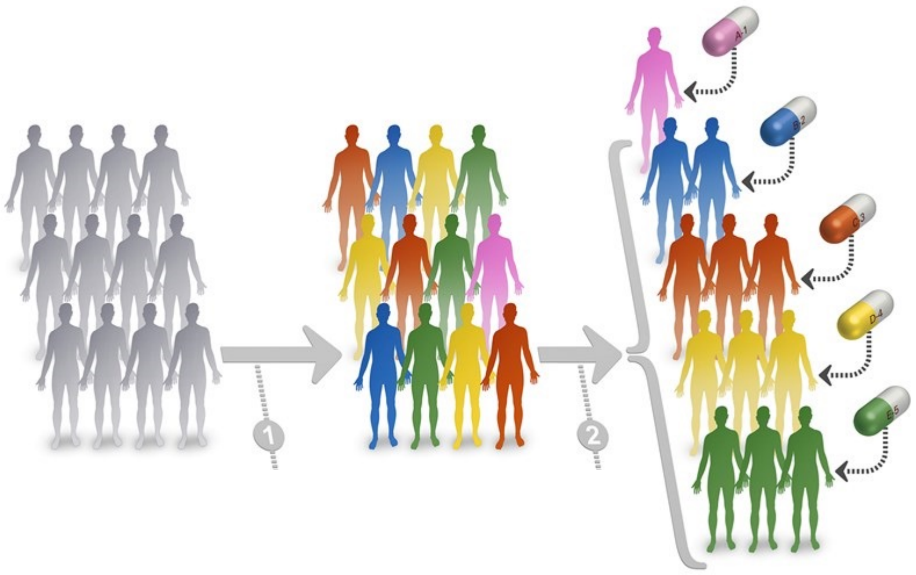
When and How Can Machine Learning be Used for Treatment Recommendations in a Clinical Setting?

CHAIR-SU Workshop: The Learning Hospital
March 2023

Uri Shalit, Technion

Using patient data to personalize treatment

- One of the ultimate promises of big data in healthcare
- Especially important when there are no clear clinical guidelines



You have condition A. Treatment options are either T=0 or T=1

I'm not so sure what's best…

Recommend T=0

Using patient data to personalize treatment: causal inference

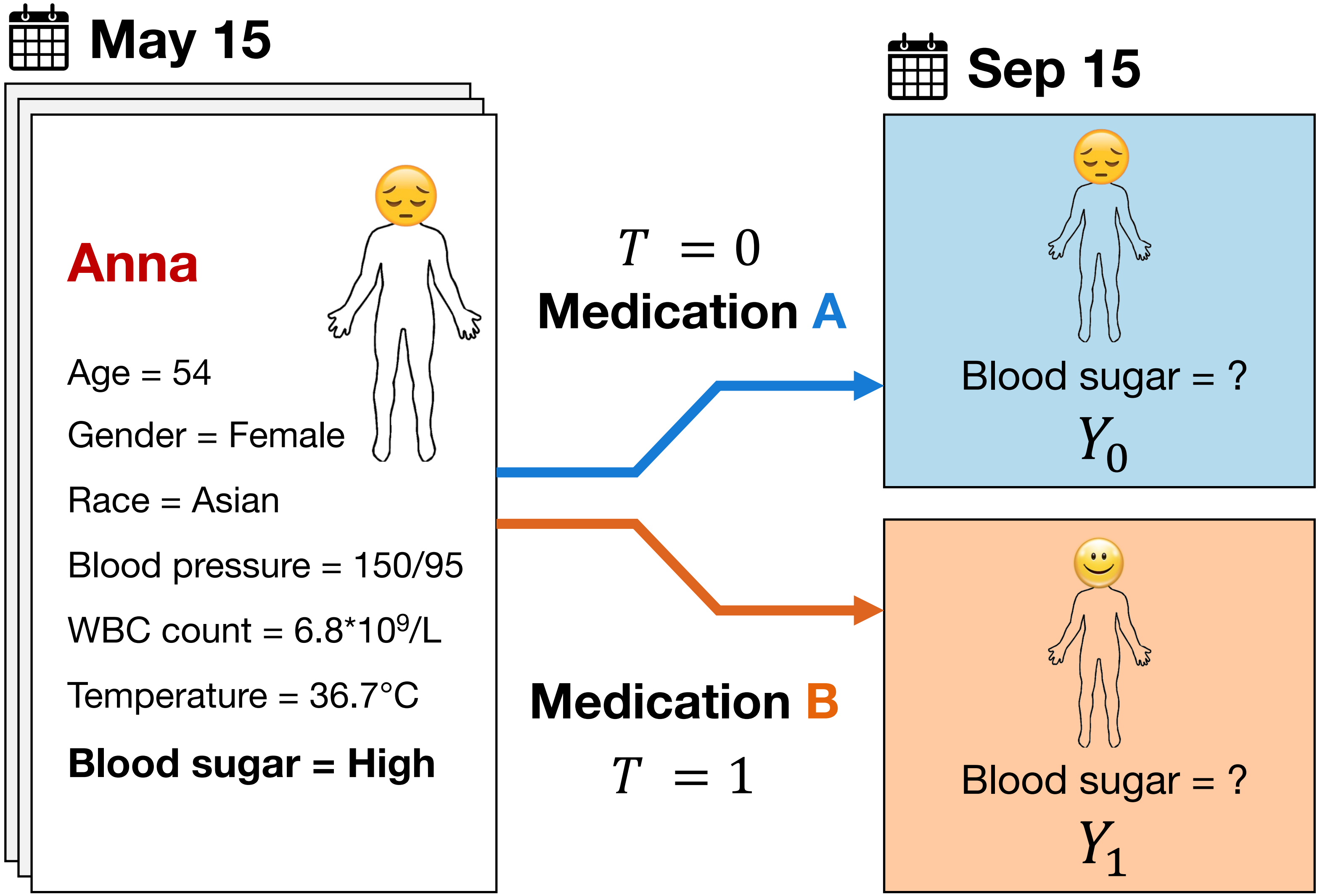
- Decision making requires **causal modeling**:
  Taking actions in the world
- Especially if model uses observational data
  - E.g. data collected from hospitals, clinics, and by patients themselves
  - Such data generally suffers from **confounding**
- No way to know if we are correct before deploying the system!
- How do we build confidence and avoid harm?

This talk

- We propose a best-practices framework for using patient clinical data to build a treatment recommendation model
  - Responsibly
  - Not focused on a specific algorithm
- Three phases:
  1. **Identification: can the data even do what I want it to do for me?**
  2. **Estimation: what does the data tell me to do?**
  3. **Validation: how much should I believe the model I just estimated?**

Y_0, Y_1: potential outcomes

(Rubin, Neyman)



May 15

Anna

Age = 54

Gender = Female

Race = Asian

Blood pressure = 150/95

WBC count = 6.8*10⁹/L

Temperature = 36.7°C

Blood sugar = High

$T = 0$

Medication A

Medication B

$T = 1$

Sep 15

Blood sugar = ?

Y_0

Blood sugar = ?

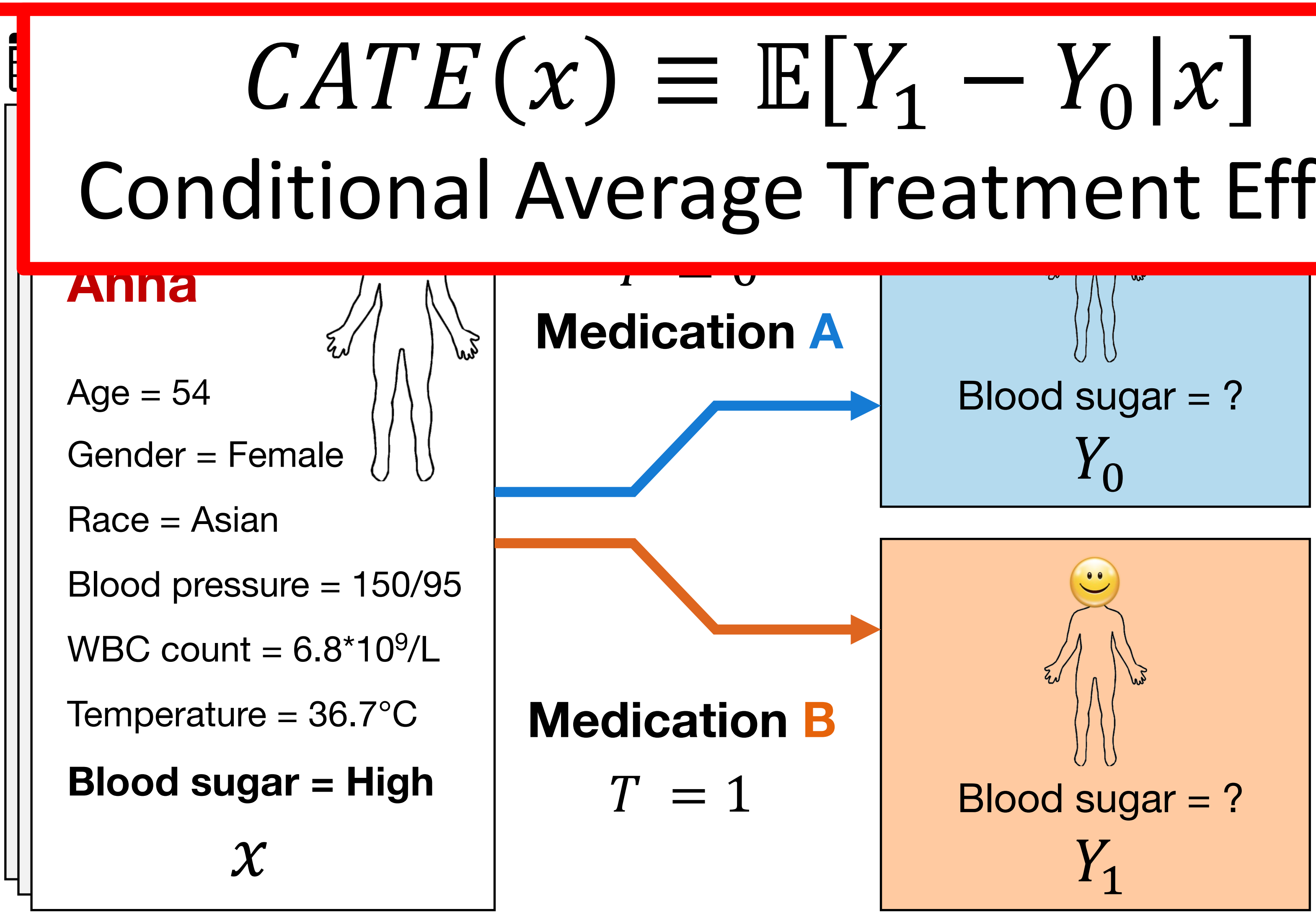
Y_1

Y_0, Y_1: potential outcomes

(Rubin, Neyman)

$$CATE(x) \equiv \mathbb{E}[Y_1 - Y_0 | x]$$

Conditional Average Treatment Effect



Anna

Age = 54

Gender = Female

Race = Asian

Blood pressure = 150/95

WBC count = $6.8*10^9$/L

Temperature = 36.7°C

Blood sugar = High

x

$T = 0$

Medication A

Blood sugar = ?

Y_0

Medication B

$T = 1$

Blood sugar = ?

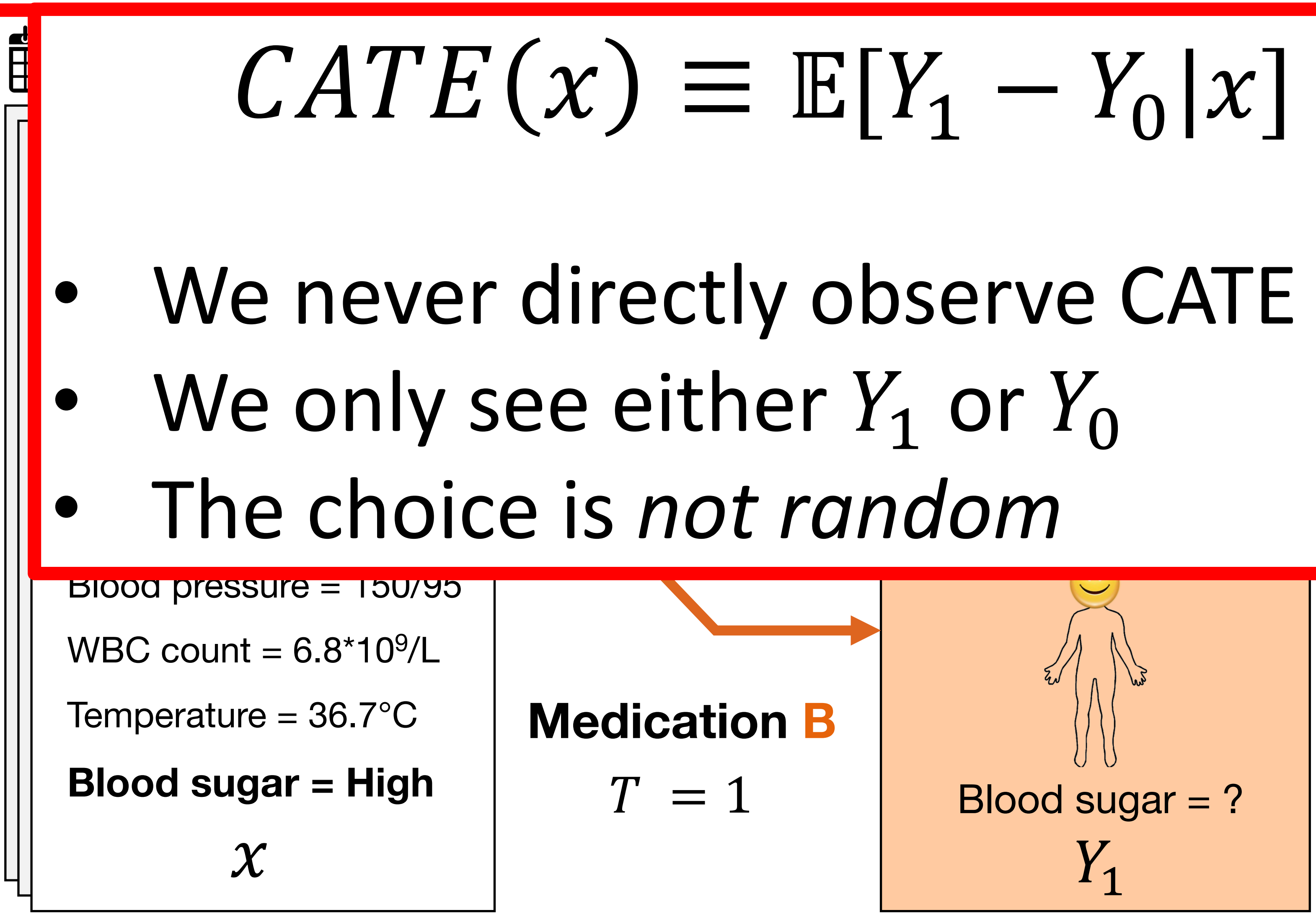
Y_1

Y_0, Y_1: potential outcomes

(Rubin, Neyman)

$$CATE(x) \equiv \mathbb{E}[Y_1 - Y_0 | x]$$

- We never directly observe CATE
- We only see either Y_1 or Y_0
- The choice is *not random*



Blood pressure = 150/95

WBC count = 6.8*10^9/L

Temperature = 36.7°C

Blood sugar = High

x

Medication B

$T = 1$

Blood sugar = ?

Y_1

Individual-level treatment effects: CATE

- We wish to estimate the individual-level treatment effect, formally denoted Conditional Average Treatment Effect (CATE)
- In Rubin-Neyman potential outcome notation:

$$CATE(x) \equiv \mathbb{E}[Y_1 - Y_0 | x] = \mathbb{E}[Y_1 | x] - \mathbb{E}[Y_0 | x]$$

"**what if** we forced the patients with features x to receive treatment $T = 1$, vs. forced them to receive treatment $T = 0$"

- We never directly observe $CATE(x)$
- We can't provably know "**what if**"

From CATE to recommendation

- $CATE(x) \equiv \mathbb{E}[Y_1 - Y_0 | x]$
- General idea:
  Estimate $\widehat{CATE}(x)$ for incoming patient with features x
- Present recommendation to doctor:
- $\widehat{CATE}(x) < 0$ → recommend $T = 1$
  $\widehat{CATE}(x) > 0$ → recommend $T = 0$

From CATE to recommendation

- $CATE(x) \equiv \mathbb{E}[Y_1 - Y_0 | x]$
- General idea:
  Estimate $\widehat{CATE}(x)$ for incoming patient with features x
- Present recommendation to doctor:
- $\widehat{CATE}(x) < 0$ → recommend $T = 1$
  $\widehat{CATE}(x) > 0$ → recommend $T = 0$
- If uncertainty about $\widehat{CATE}(x)$ is high→ **defer** recommendation

Individual-level treatment effects: CATE

- $CATE(x) \equiv \mathbb{E}[Y_1 - Y_0|x]$
- x is high-dimensional and practically unique to each unit
- Can (carefully) use machine learning based tools
  - Causal Forests (Wager & Athey 2015, 2018), Deep networks (Johansson, S, Sontag 2016, 2017, Parbhoo et al. 2018, Shi et al. 2019), Gaussian processes (Schulam & Saria 2018, Alaa & van der Schaar 2018), Meta-learning (Künzel et al. 2017, 2019, Nie & Wager 2017)
- However: These only work *under a strict set of* ***causal identification*** *conditions:*
  - no hidden confounders
  - common support between different treatments
  - no interference between units
- Most of these assumptions are not testable from data
- (Even supervised learning will not work unless the conditions hold)

Major challenges

1

- All causal effect estimation methods rely on causal identification assumptions, some of whic
- Prominently: **re measured”**
- Some typical l
  - Treatment a
  - Treatment a ne
  - Important c nodel, e.g. imaging

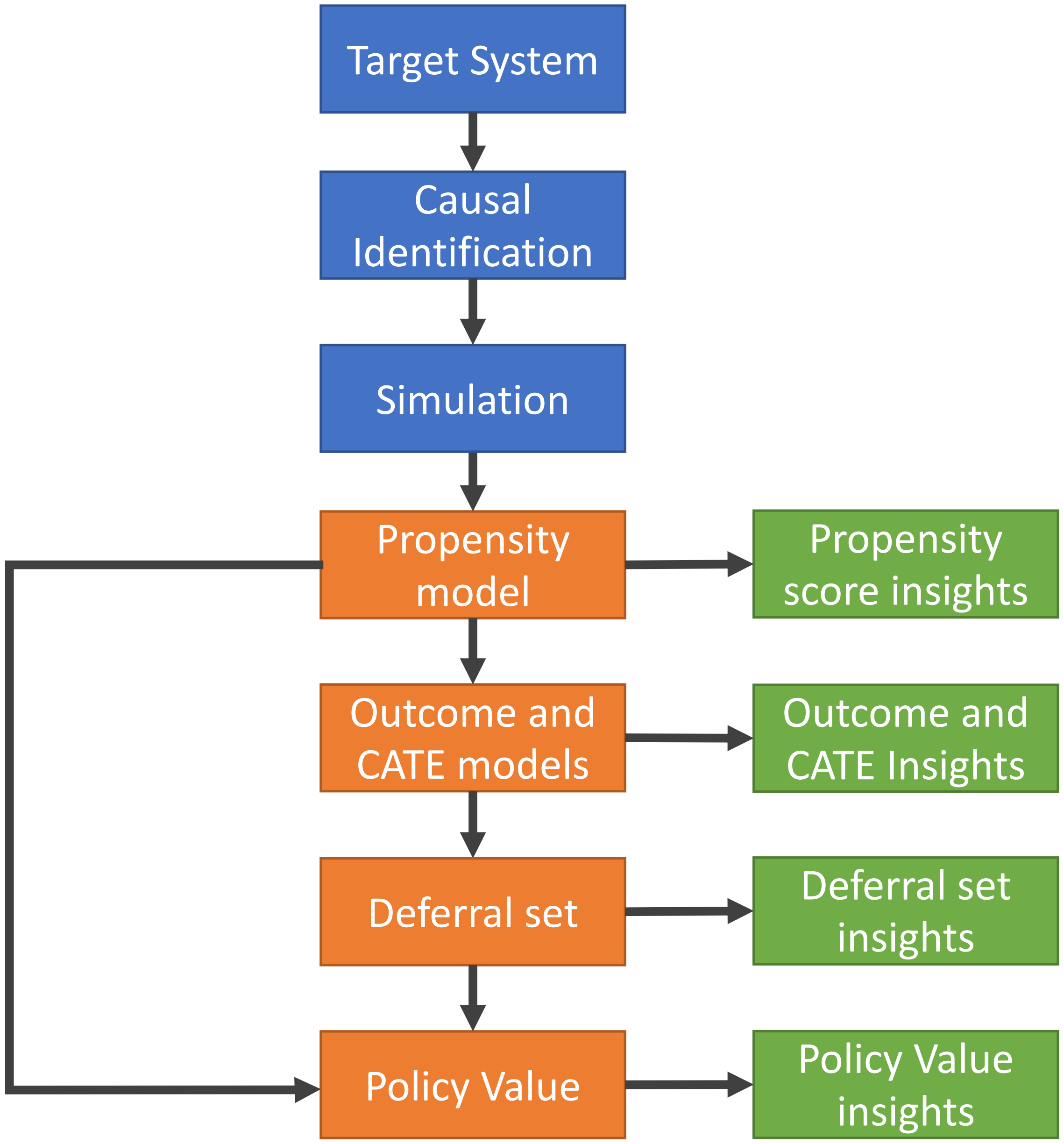
How can we still build confidence and deploy a treatment recommendation system ?

2

- **There is no test set**
  - When our recommendation differs from what happened in practice → can’t know for sure what would have happened had recommendation been used

- High stakes: even a pilot system might cause harm

Framework for robustly building causal decision support models



Target System

Causal Identification

Simulation

Propensity model

Propensity score insights

Outcome and CATE models

Outcome and CATE Insights

Deferral set

Deferral set insights

Policy Value

Policy Value insights

🟦 ***Identification***

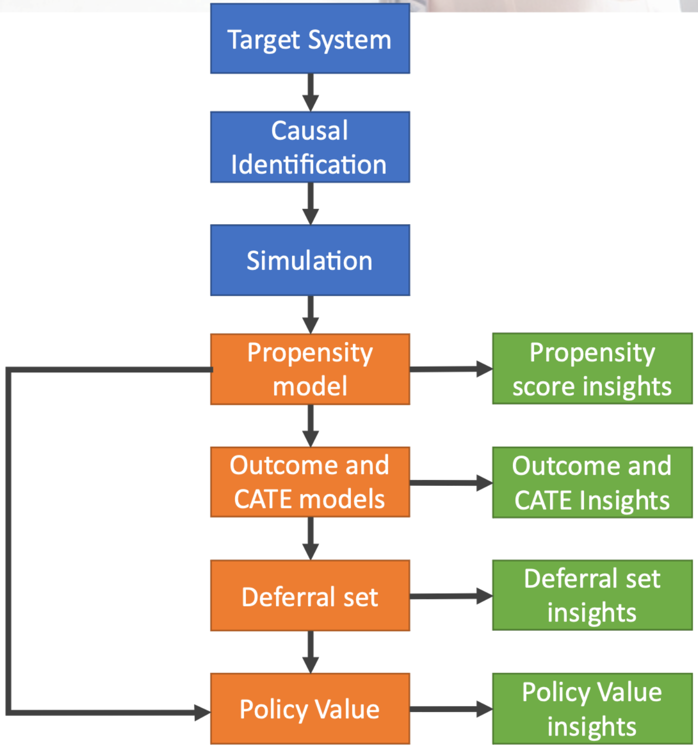
🟧 ***Estimation***

🟩 ***Validation***

Identification I: The Target System

(following Miguel Hernán's "Target Trial")

Define exactly the setting and context of the treatment recommendation system



Identification I: The Target System

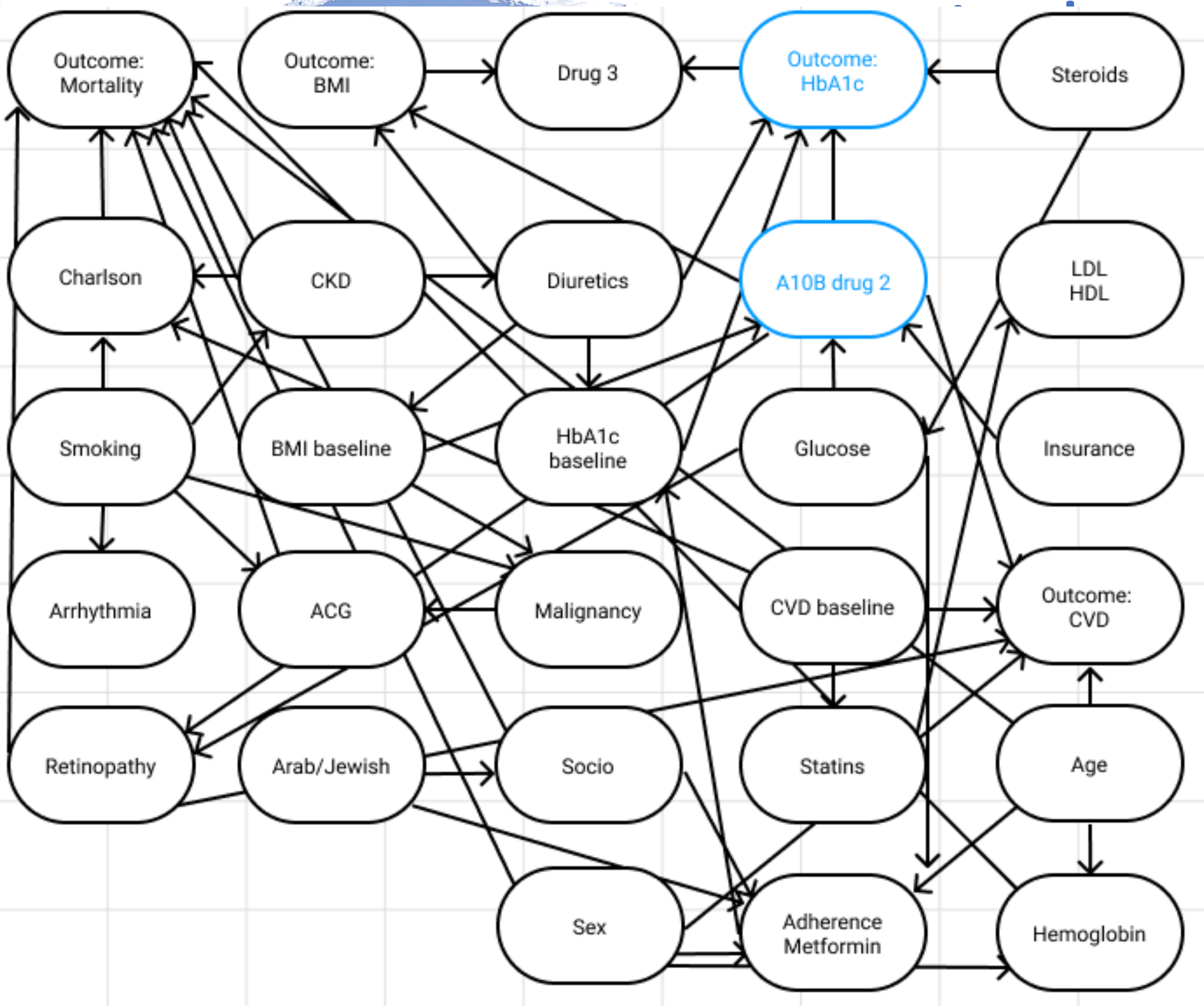
(following Miguel Hernán's "Target Trial")

Points for discussion with clinical partners:

1. Is treatment decision made by physicians at a **well-defined point in time?**
2. Is the set of possible actions small?
3. Are there clear clinical guidelines for decision?
4. Is there high variability in treatment decisions between physicians?
5. Are there well-defined and widely agreed upon outcomes?

Help clarify discussion with clinicians about "AI assistants"

Identification II:
data suitable for
ng the target system?



- For observational data, have we measured all (most) known confounders?
- Do we have temporal separation of what data is recorded before/after the treatment assignment?
- Causal graphs built with domain experts can be useful here

First return point: no identification

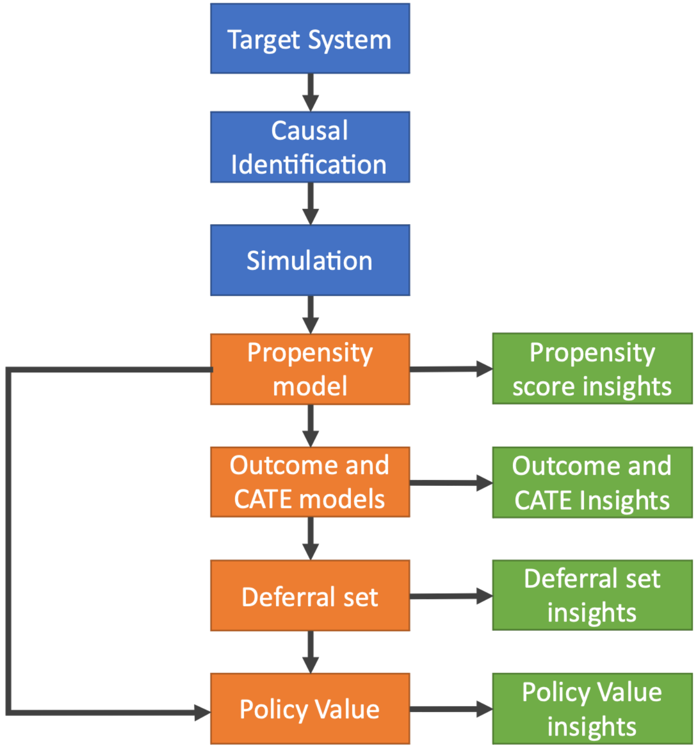
First return point: no identification

Estimation I + II: CATE and propensity score

Many great methods for estimating:

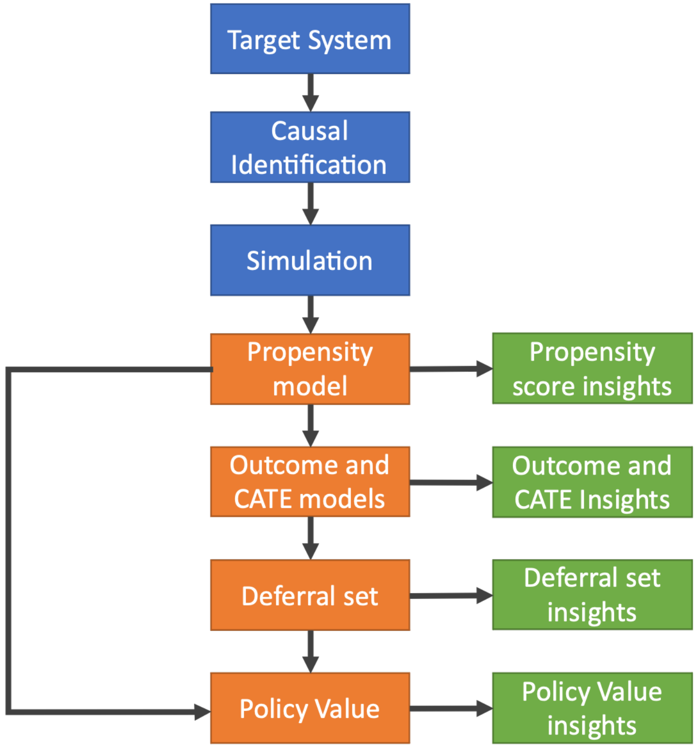
- Propensity scores $p(t = 1|x)$
- $CATE(x)$

Will not go into details here!



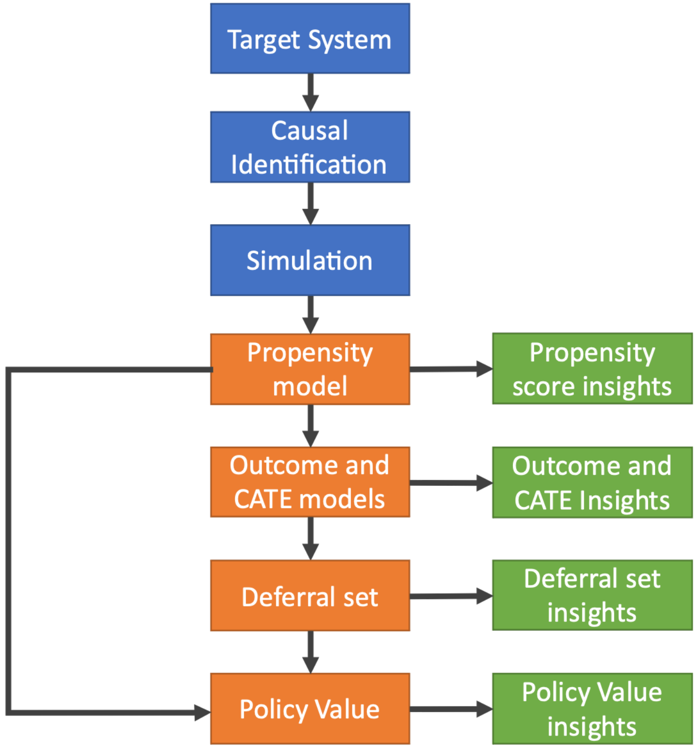
Validation I + II: CATE and propensity score

- No traditional test set for CATE
- Still: should evaluate regression/classificaiton models with standard metrics (MSE, AUC, PPV, etc.)
- Compare diff y
- Apply interpr
- Check with c
- Characterize
  - Positive and
    Who do we recommend receive each of the treatments?
  - Clinician aggrement and disagreement sets:
    Where do we think the clinicians were wrong?



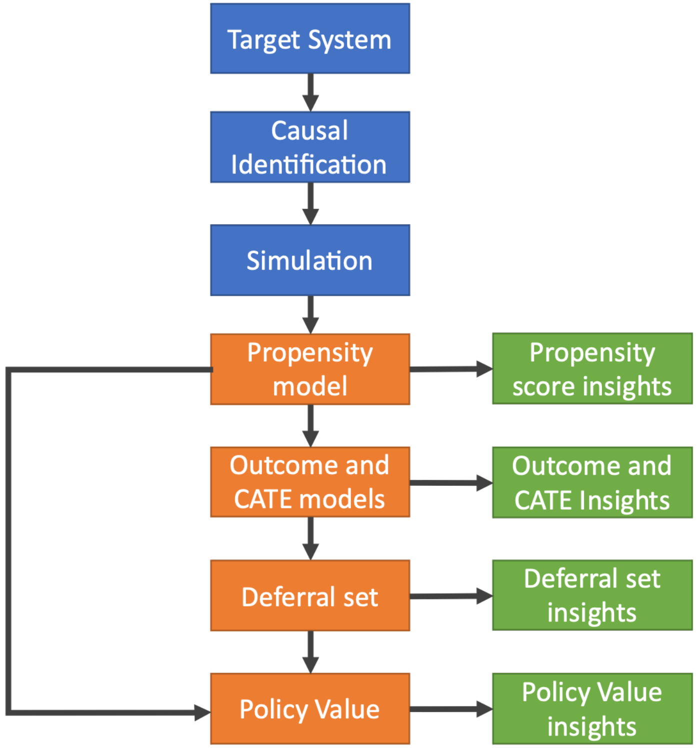
Estimation III: Deferral Set

- Why we might doubt a given $CATE(x)$ estimate:
  - Estimation error
    (finite samples, weak overlap, model mis-specified, covariate shift)
  - Noise (measurment error, outcome stochasticity)
  - Causal error (hidden confounding)
- If in doubt, might wish to defer decision
- We work on modeling all sources of error



Estimation IV: Policy Value

- Crucial metric for recommendations:
  "What would be the expected outcome if physicians treated as the model recommends"
- The **policy value** of the current "doctors' policy" is simply the average outcome in the population
- A good recommendation would have a better policy value than the doctors' policy
- Estimating policy values is itself a challenging causal problem
  - Don't really know what would have happened in cases where our recommendations differ from the actual treatment in the data



Preliminary results

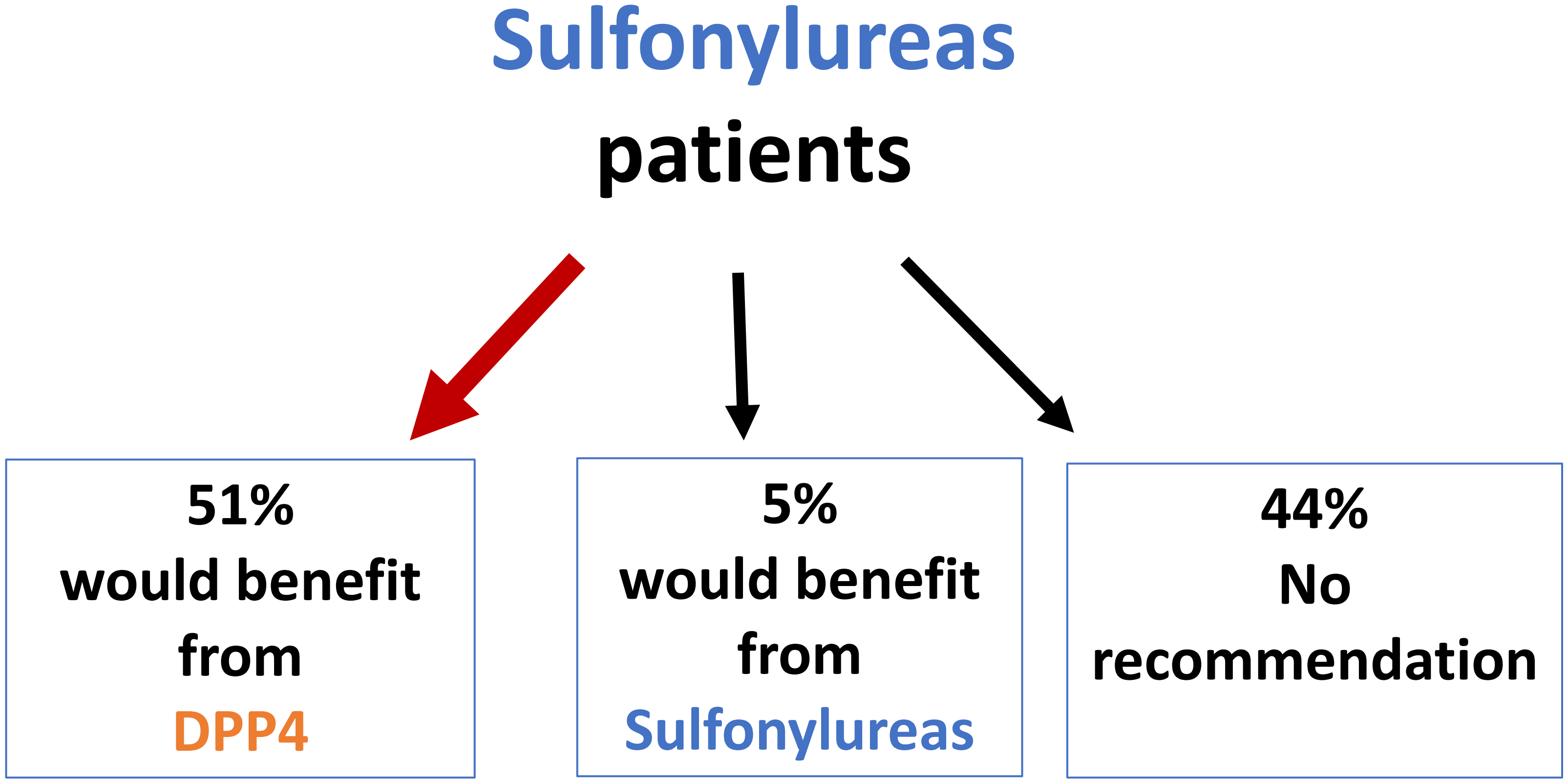
- Analysis lead by graduate students Rom Gutman (Technion) and Shimon Sheiba (Technion, Cytoreason)

Preliminary results 1: Chronic disease

Clalit Research Institute

- Joint work with Clalit Research Institute (Prof. Ran Balicer)
- Investigating the effects of Sulfonylureas vs. DPP4 on type-II diabetes patients who have not responded to first-line therapy
- Goal: reduce blood sugar, measured in A1C
- More than 50,000 patients
- More than 200 covariates which are potential confounders: demographics, lab tests, diagnoses, medications, administrative and more

Preliminary results – optimal care



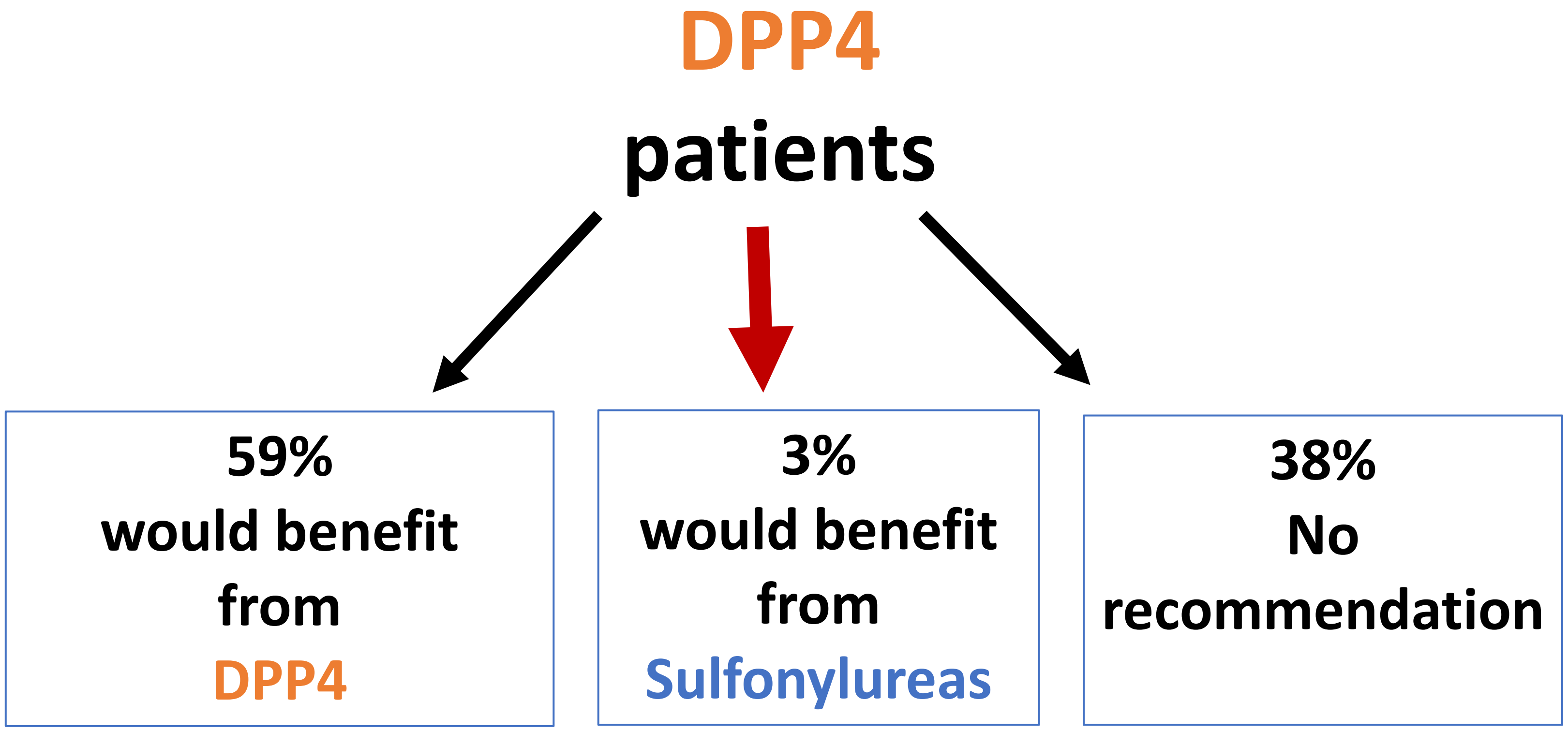
Sulfonylureas
patients

51%
would benefit from
DPP4

5%
would benefit from
Sulfonylureas

44%
No recommendation

Preliminary results – optimal care



DPP4 patients

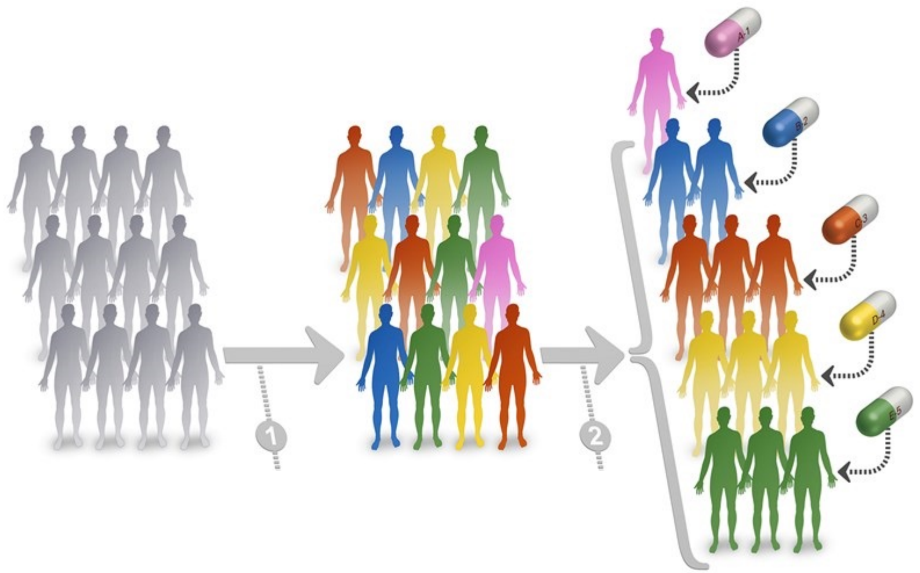
59% would benefit from DPP4

3% would benefit from Sulfonylureas

38% No recommendation

Is personalization worth it?

- Between Sulfonylureas and DPP4, the answer: no!
- We detect no significant differenece between:
  a) moving *everyone* to DPP4
  b) personalized treatment
- Clinical trials later showed advantage of DPP4
- Newer medications are now in use



Preliminary results: Acute disease

- The causal effects of diuretics on kidney function in hospitalized acute heart failure patients with kidney injury in Rambam Medical Center
- Clinical collaborators:
  Dr. Oren Caspi and Prof. Doron Aronson
  (Technion University & Rambam Health Care Campus)

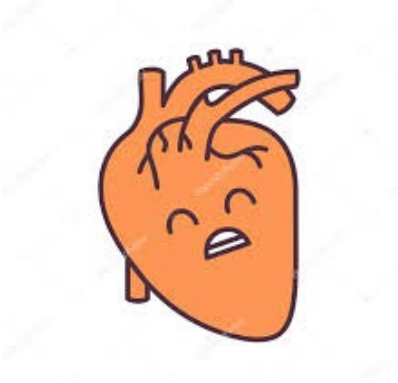
Preliminary results:
Heart failure with kidney injury

RAMBAM

- Causal effects of *diuretics* on *kidney function* in hospitalized acute heart failure patients with kidney injury in Rambam Medical Center
- Physicians tell us:
  They have poor guidance how to prescribe diuretics and blood-pressure medications to these patients
- 2157 hospitalized heart failure patients with rise in serum creatinine, indicating kidney injury
- More than 200 covariates which are potential confounders:
  demographics, lab tests, diagnoses, medications, administrative and more
- Empirically: half of cohort had increased diuretics or leave the same,
  half had decreased diuretics

Preliminary results
Heart Failure with kidney injury

- T=1: “Decrease diuretics”
  - Often improves kidney function
  - Might hurt cardiac function
- Must balance multiple outcomes
- *“Should we increase, keep or decrease diuretics for this patient?”*



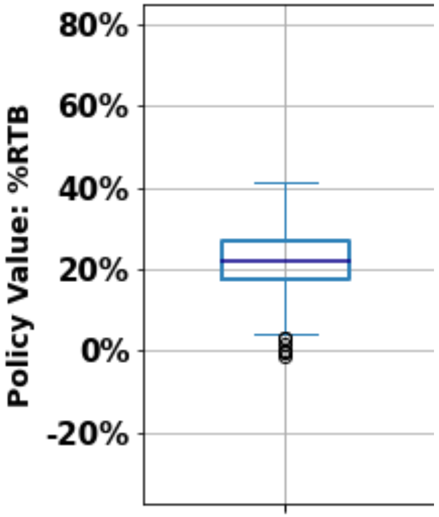
RESULTS: Held-out cohort

(n=530)

Compare outcomes under current practice vs. proposed Causal Machine Learning Model

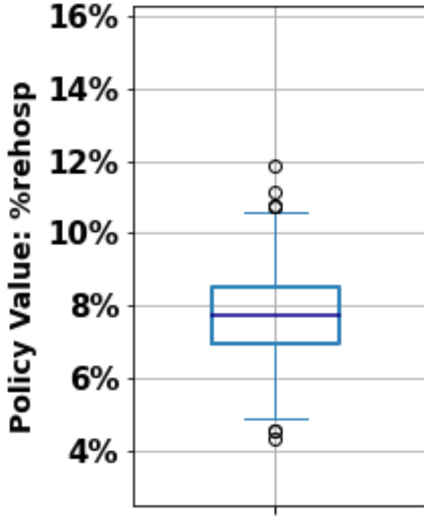
Current practice

Top: kidney function (higher=better)
%RTB = %Return-to-baseline creatinine



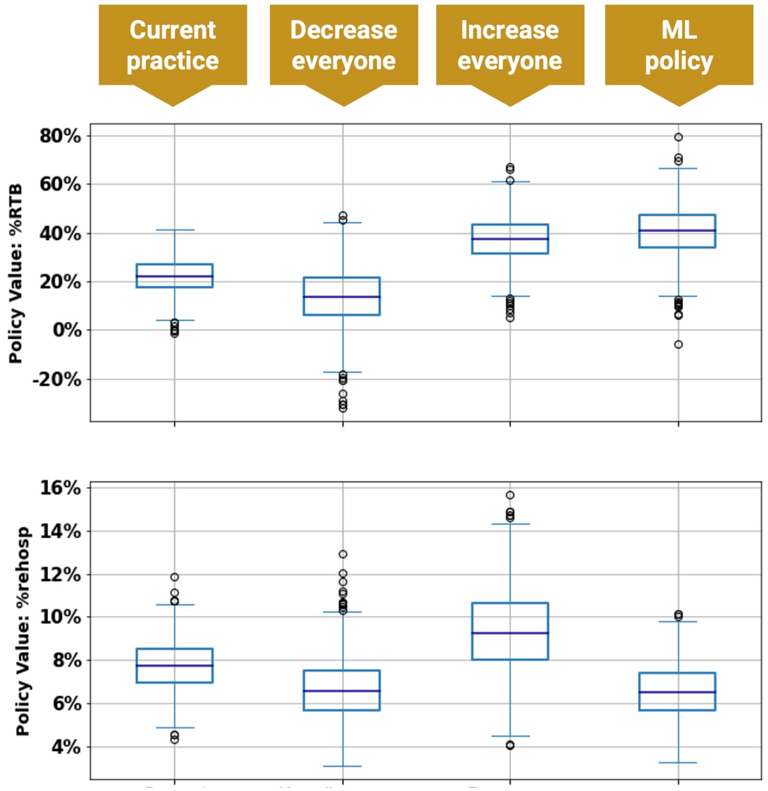
Bottom: rehospitalization
(indicator of cardiac function, lower=better)

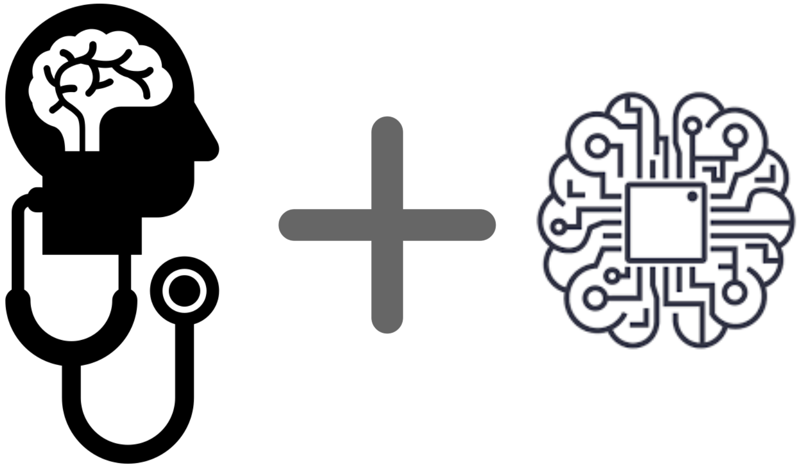
- Our recommendations are better than current practice for %RTB (p=0.015, median 41% vs. 22%) and somewhat better for rehosp. (p=0.048, median 6.5% vs. 7.7%)



Should we believe this?

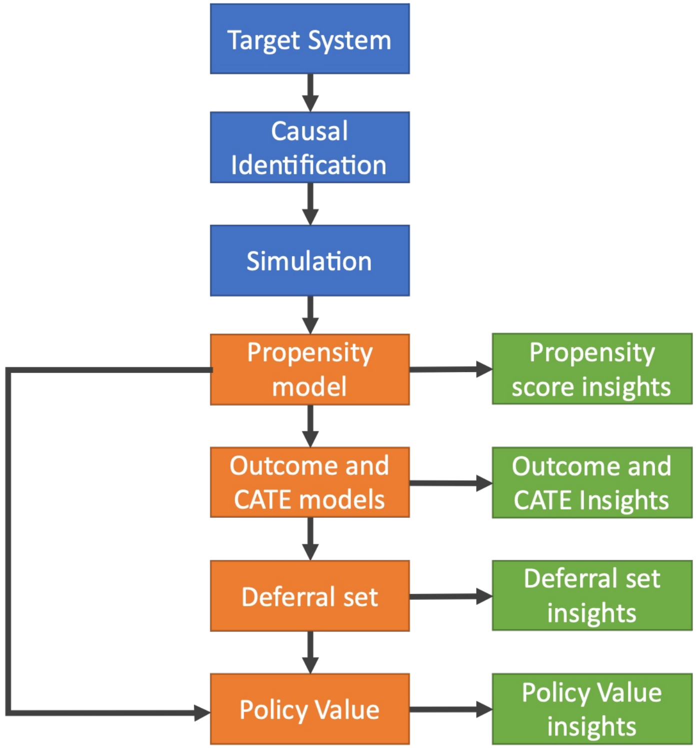
- We don’t **really** know what would have happened if our recommendations would have been followed
- The only way to truly know if our recommendations are useful: Run an experiment testing the system with real physicians and patients

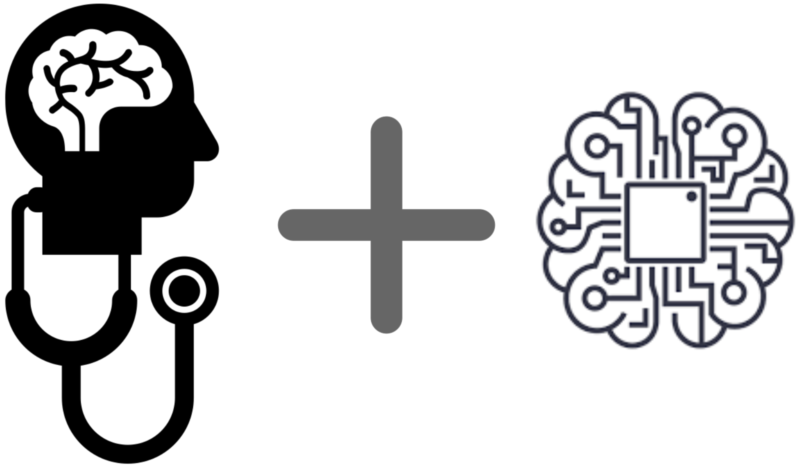




Should we believe this?

- We don’t **really** know what would have happened if our recommendations would have been followed
- The only way to truly know if our recommendations are useful: Run an experiment testing the system with real physicians and patients
- The goal of our framework is to come in the best possible safe shape towards such a trial
- **Currently planning the trial**





Planning an “AI@hospital trial”

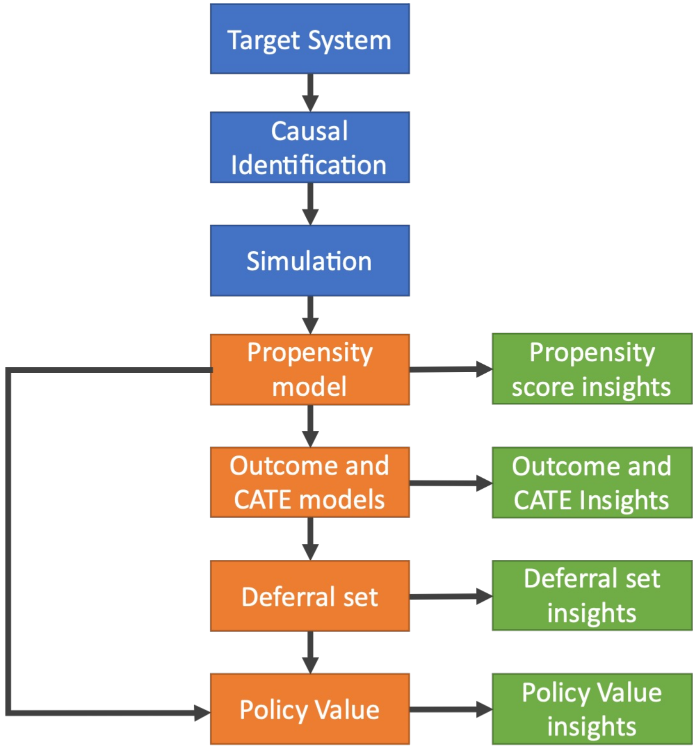
- Rambam Healthcare Campus:
  - 1,000 bed tertiary hospital
  - serving a population of 2,000,000 people
- TERA: Technion-Rambam Initiative in Medical AI
  - Jointly funded center
  - Money, clinician time, and space for joint research
  - Expedited access to data (regulatory and technical)
  - Support for deployment at bedside
- In practice
  - Strong clinical-computational collaboration on a personal level is key
  - Joint commitment to goal and willingness to invest time & energy in it (especially clinician’s time!)

TERA

Technion-Rambam Initiative in Medical AI

Planning an “AI@hospital trial”

- *Now funded for trial*
- Plan:
  - Intense discussions with entire clinical team
  - Deepen understanding of clinical workflow
  - Where does the system come in?
  - Push or pull?
  - Unit of Randomization?
    - “Defer” or “Run Model”
- Main outcomes
  - Clinician acceptance, adherence, reaction
  - Safety
  - Kidney function
- **Looking forward to your thoughts and comments!**



Target System

Causal Identification

Simulation

Propensity model

Propensity score insights

Outcome and CATE models

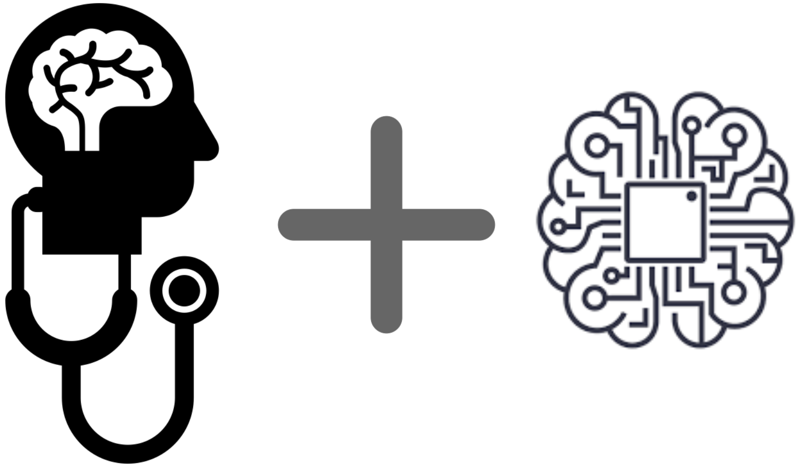
Outcome and CATE Insights

Deferral set

Deferral set insights

Policy Value

Policy Value insights



We are hiring!

- Joint PhD / Postdoc with Prof. Mihaela van der Schaar at Cambridge University
- Work on methods for causal inference and machine learning in healthcare
- Email shalit-lab@technion.ac.il

TECHNION
Israel Institute
of Technology

UNIVERSITY OF
CAMBRIDGE

Thank you!

- Technion:
  - Rom Gutman
  - Shimon Sheiba
  - Omer Noy
- Rambam Health Care Campus & Technion:
  - Dr. Oren Caspi
  - Prof. Doron Aronson
- Clalit Research Institute:
  - Ohad Levinkron
  - Dr. Janni Yuval
  - Galit Shaham
  - Dr. Becca Feldman
  - Prof. Ran Balicer