THE 'DETECTING ALGORITHMIC BIAS' (DAB) PIPELINE

CPDD 2025, New Orleans

Gabriel J. Odom, Aaron Marker, Ganesh Jainarain, Sal Giorgi, Clinton Castro, Larry Au, H. Andrew Schwartz, and Laura Brandt

DISCLOSURES

Authors have no conflicts of interest to declare.

This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

LINK TO SLIDES

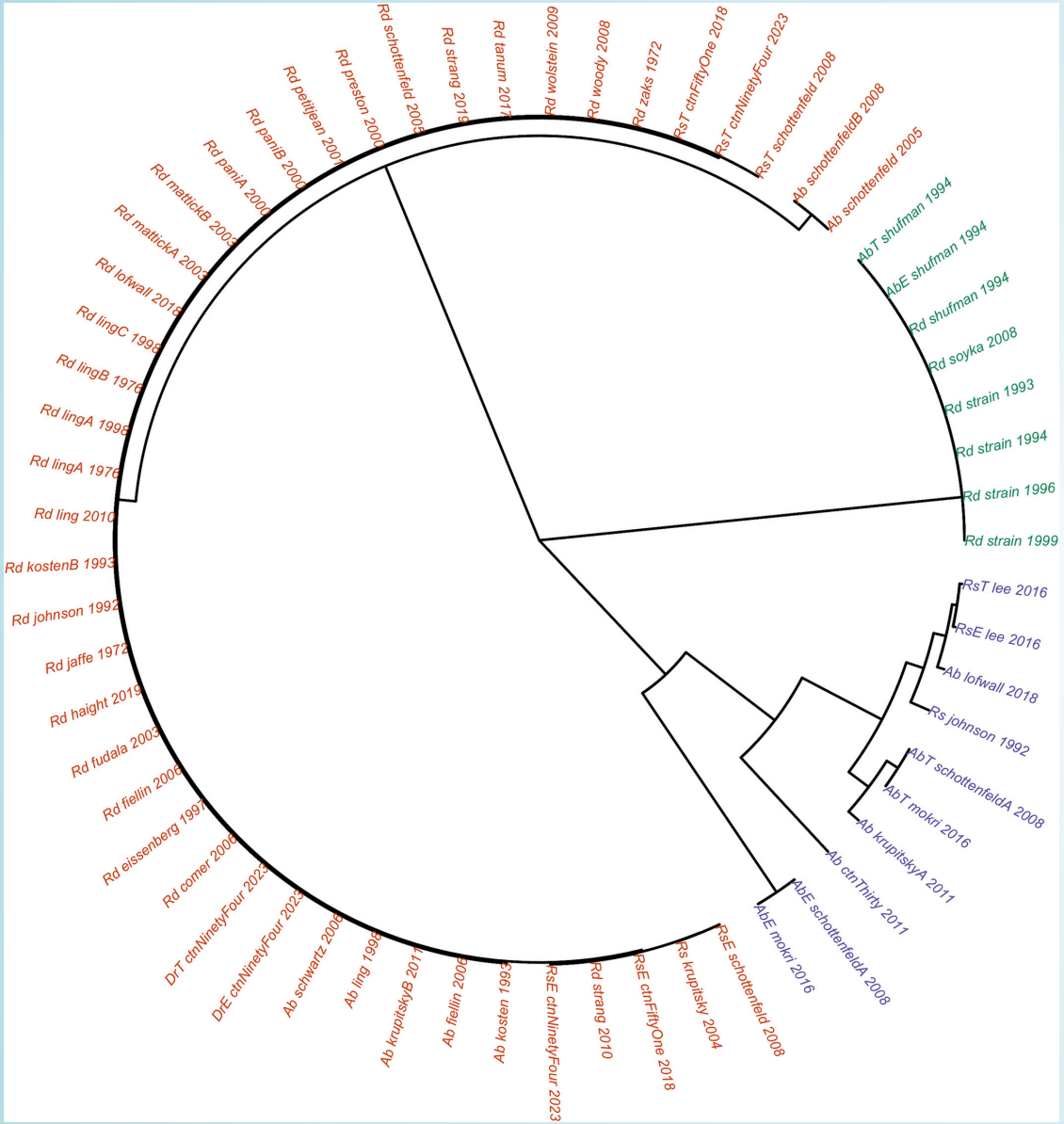
OUTLINE

1. What's the (End)Point?
2. Theoretical Foundations:
   - Stress-Testing AI/ML Models
   - What is a "Synthetic Clinical Trial"?
3. What can DAB do for You?
   - DAB Pipeline Steps
   - Real Data Example
4. Interpreting Results for…
   - Researchers
   - Patients and Care Providers

WHAT'S THE POINT?

WHAT'S THE POINT?

- An endpoint is a "targeted outcome of a clinical trial that is statistically analyzed to help determine the efficacy and safety of the therapy being studied"
- **Endpoint Science** is the study of those endpoints, especially from an empirical/rigorous perspective
- The **DAB Pipeline** enables us to study how endpoints (*and the AI/ML models predicting them*) "react" to changes in the underlying data



BUT WHY?

As Brandt, Odom, et al. (2024) showed:

- Not all endpoints are created equal
- Endpoints don't always measure what they claim

GARBAGE IN—GARBAGE OUT

- Artificial Intelligence and Machine Learning (AI/ML) models offer the potential for profound translational insights in SUD treatment, *if they are used correctly*
- AI/ML models are limited to the data you give them: these models will do their best to predict the endpoint you choose based on the demographic and clinical predictors you supply
- Unless you choose your endpoint carefully, AI/ML will give you the **right answer** to the **wrong question**

THEORETICAL FOUNDATIONS

THEORETICAL FOUNDATIONS

- Stress-Testing AI/ML Models
- What is a “Synthetic Clinical Trial”?

STRESS-TESTING MODELS

- **Stress testing** in engineering involves taking some constructed tool or product and “pushing” the product to its limits (and beyond)
- Products are stress tested to ensure they are safe and effective to use
- AI/ML models are often built in “clean” research environments, free of many bizarre components of human behavior. As such, they need stress testing before they can be applied to new data.

EXAMPLE: SELF-DRIVING CARS

Could you force a self-driving car to take “Exit 7 for Bloomingdale Rd”?

Can you trap a self-driving car with a salt circle? With cones?

STRESS TESTING ENDPOINTS

- Stress testing predictive models is good, but can still suffer from the “right answer to the wrong question” problem
- The DAB pipeline is designed to stress test clinical trial endpoints which have been used to evaluate medication-based treatments for SUD
- We want to use clinical trial endpoints that have good resilience and stability across various “stress” conditions

SYNTHETIC CLINICAL TRIALS

- Endpoints are applied to their respective clinical trials, but will they have similar performance in a new trial?
- A **synthetic trial** is a new data set created in the computer from previously published clinical trial cohorts
- Because participants in synthetic cohorts are "recruited" from clinical trials or EMR databases, we get two benefits: 1) thousands of cohorts can be created instantly, and 2) these cohorts can have "bespoke" demographics and patient characteristics

FINE-GRAIN COHORT CONTROL

With Synthetic Cohorts, we can create random trial cohorts from real data that represent hard-to-recruit subpopulations. For example, cohorts could have:

- 0% - 100% racial/ethnic minority participation
- 0% - 100% female participation
- 0% - 100% rural participation
- only people who inject drugs
- overrepresentation of a specific age group
- and many more

WHAT CAN DAB DO FOR YOU?

CONTROLLED STRESS TESTING WITH THE DAB PIPELINE

- Each of the synthetic cohort “stresses” the endpoint (and AI/ML model predicting it) in various ways
- We perturb cohort characteristics across various percentage values to change how the endpoint and its model will be tested
- The DAB Pipeline generates 1000s of these synthetic clinical trial cohorts (under potential “stressors”), evaluates the endpoint and model across these conditions, and returns evaluation data

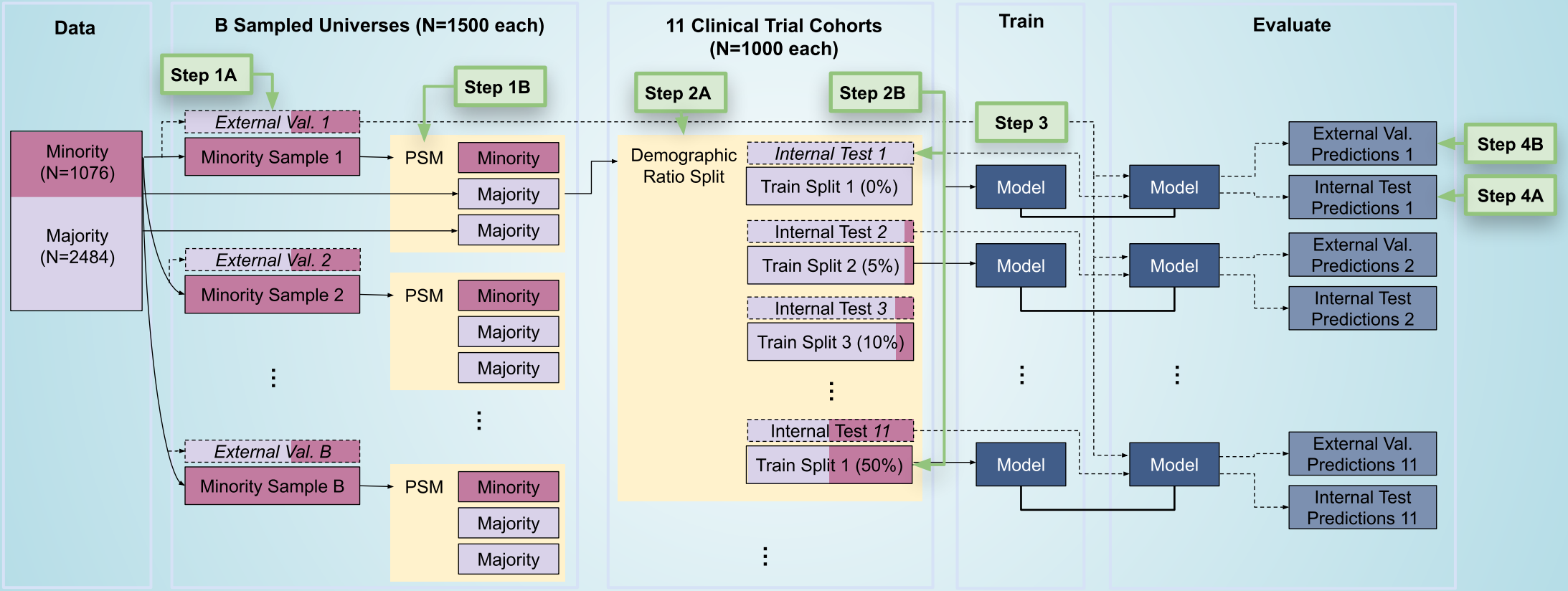
PIPELINE EXPERIMENTAL INPUTS

- A rich database of SUD patients, as a clean model data matrix (defaults to the CTN-0094 Database)
- Clinical endpoint calculated for each patient (defaults include abstinence, relapse, and use reduction endpoints from the CTNote Library)
- AI/ML model to predict endpoint (defaults to the LASSO Logistic SUD model from Luo et al. (2024))
- A binary subset “stressor” on the endpoint and model. As an example, we will use increasing minority representation in the synthetic cohorts

NON-DEFAULT PIPELINE ARGUMENTS

```
python run_pipelineV2.py --loop 1 11 --col is_minority --samp 500 --dir
results
```

DAB Pipeline



PIPELINE STEPS (PT 1)

1. a. *External Validation Data*: sample 58 non-Hispanic white and 42 minority participants

   b. *Train-Test Universe*: sample `--samp 500` minority participants and use propensity score matching to sample 2 sets of 500 matched non-Hispanic white participants

2. a. create 11 synthetic clinical cohorts with $\mathbf{n} = 1000$, with minority representation ranging from 0% - 50%

   b. create 75-25 train-test splits for each of the 11 synthetic clinical trial cohorts in 2a

PIPELINE STEPS (PT 2)

PIPELINE OUTPUTS

The DAB Pipeline returns results in subdirectories on your computer:

- Model predictions and evaluations in external validation data (`heldout_predictions/` and `heldout_evaluations/`)
- *For model checking only:* Model predictions and evaluations in testing data (`subset_predictions/` and `subset_evaluations/`)
- Logging messages and error statements (`logs/`)

PERFORMANCE EVALUATION

Within the `heldout_evaluations/` directory (validation results), there will be one `.csv` file per random seed value and endpoint with columns for ...

- *Meta-data*: the global seed value, endpoint type, endpoint name, and AI/ML model script name
- *Sample counts for selected "stressor" feature*: name of demographic comparison feature, sample sizes in prediction data, and sample sizes in training data
- *Performance metrics*: depends on the endpoint type, but could include accuracy, F1, R^2, pseudo-R^2, etc.

EVALUATION DATA (PT 1)

global_seed	Outcome Type	Outcome Name	Model script name	Demog Comparison	Prop(Demog)	Training Demographics
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	750 1
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	712 1, 21 3, 11 2, 6 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	668 1, 43 3, 20 2, 19 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	632 1, 61 3, 31 2, 26 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	612 1, 71 3, 34 2, 33 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	558 1, 100 3, 51 2, 41 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	516 1, 121 3, 57 2, 56 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	473 1, 141 3, 75 2, 61 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	445 1, 159 3, 84 2, 62 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	410 1, 181 3, 88 2, 71 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	376 1, 191 3, 111 2, 72 4

For these results, the numeric **seed** was 0, the **endpoint type** was binary (success or failure), the **endpoint** was an *abstinence* measure from Krupitsky et al. (2011), and the **AI/ML model** is the default LASSO Logistic model for OUD from Luo et al. (2024).

EVALUATION DATA (PT 2)

global_seed	Outcome Type	Outcome Name	Model script name	Demog Comparison	Prop(Demog)	Training Demographics
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	750 1
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	712 1, 21 3, 11 2, 6 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	668 1, 43 3, 20 2, 19 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	632 1, 61 3, 31 2, 26 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	612 1, 71 3, 34 2, 33 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	558 1, 100 3, 51 2, 41 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	516 1, 121 3, 57 2, 56 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	473 1, 141 3, 75 2, 61 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	445 1, 159 3, 84 2, 62 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	410 1, 181 3, 88 2, 71 4
0	Binary	Ab_krupitskyA_2011	TBD	Race: non hispanic white vs minority	58 1, 19 3, 16 2, 7 4	376 1, 191 3, 111 2, 72 4

The endpoint and model **stressor** was non-Hispanic white vs minority, the **validation demographic proportion** was 58% (US demographic), and the **synthetic cohort demographic proportion** increases from 0% minority to 50% minority in 5% increments.

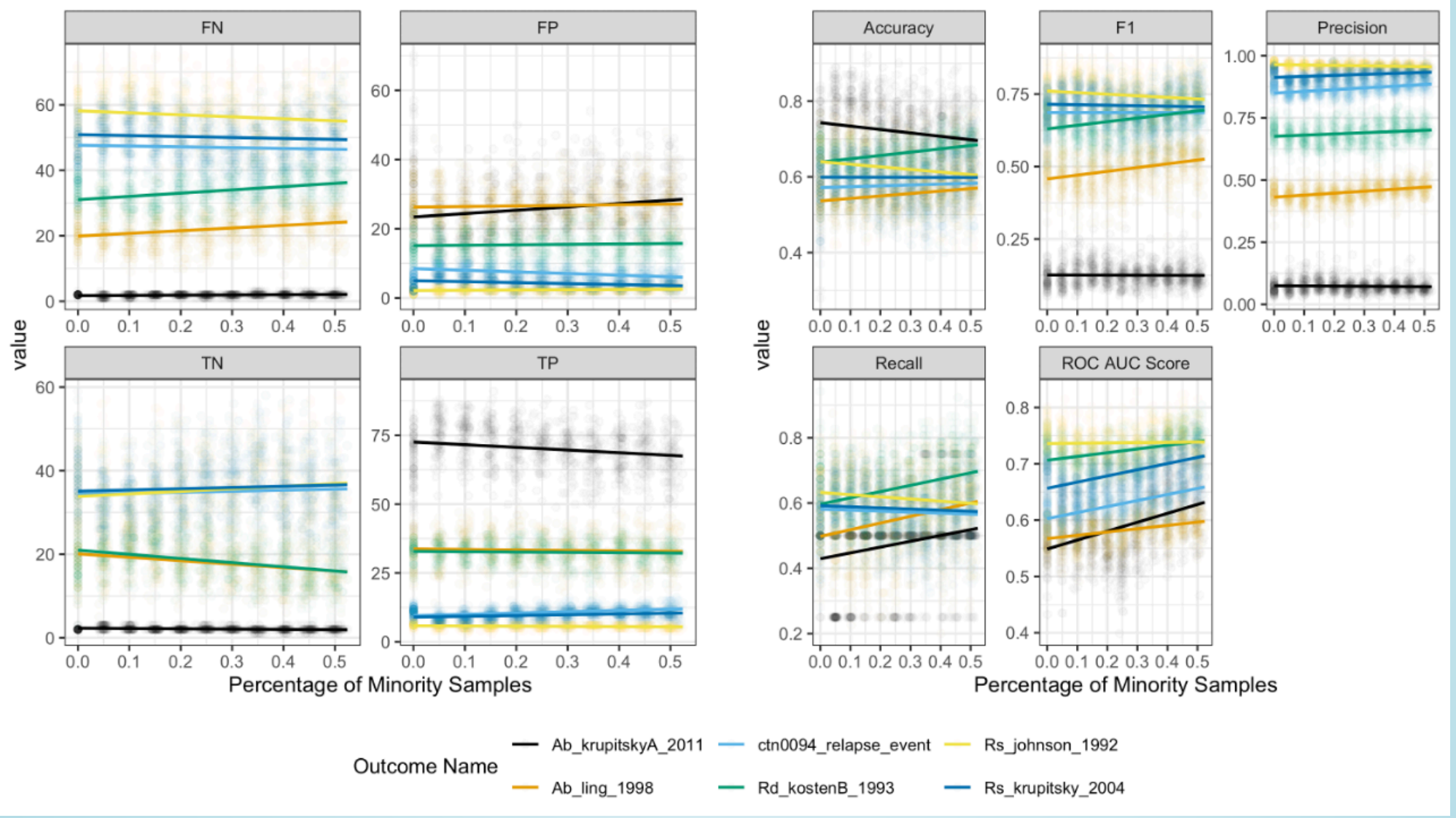
EVALUATION DATA (PT 3)

TP	TN	FP	FN	Accuracy	Precision	Recall	F1	ROC AUC Score
72	2	24	2	0.74	0.07692308	0.5	0.13333333	0.549479167
81	2	15	2	0.83	0.11764706	0.5	0.19047619	0.645833333
77	2	19	2	0.79	0.0952381	0.5	0.16	0.59375
77	2	19	2	0.79	0.0952381	0.5	0.16	0.520833333
50	2	46	2	0.52	0.04166667	0.5	0.07692308	0.5546875
77	2	19	2	0.79	0.0952381	0.5	0.16	0.5390625
57	2	39	2	0.59	0.04878049	0.5	0.08888889	0.528645833
72	2	24	2	0.74	0.07692308	0.5	0.13333333	0.666666667
74	2	22	2	0.76	0.08333333	0.5	0.14285714	0.666666667
56	1	40	3	0.59	0.06976744	0.75	0.12765957	0.661458333
65	2	31	2	0.67	0.06060606	0.5	0.10810811	0.536458333

The **performance metrics** include the Confusion Matrix, classification accuracy, and AUROC Score. These performance metrics are all appropriate to a **binary** endpoint.

INTERPRETING RESULTS

WE WANT FLAT LINES!



PRACTICAL SIGNIFICANCE

- How flat is “flat”? We don’t have a “magic” p-value here
- Which metrics matter more to clinicians?
- Which metrics matter more to patients and their advocates?

Statisticians are not the authority over the real impact of these results. These are questions that should be discussed in a group of people from diverse academic, clinical, experiential, and cultural backgrounds.

Your Computer Can Detect Patterns—But It Can't Tell You What Matters.

- **Computational audits can flag disparities**, such as:
  - Differences in model performance across racial or ethnic groups
  - Shifts in outcome classification when endpoint definitions change
  - ...but these findings raise new questions that machines can't answer

Why We Need Humans in the Loop

🔍 **Interpretation**

- Is the observed bias *clinically significant*?
- Is it *ethically problematic*, or a reflection of meaningful group differences?

⚖️ **Moral Reasoning**

- Should we change the algorithm, the model, or the clinical workflow?
- How do we weigh performance trade-offs (e.g., recall vs. fairness)?

🗣️ **Contextual Prioritization**

- Researchers may prioritize statistical validity; patients and clinicians may prioritize *trust*, *access*, and *dignity*.
- Stakeholders bring insights about harms that aren't captured in metrics.

Our approach is grounded in three interdependent yet tightly interlinked pillars—empirical audits, ethical reasoning, and stakeholder engagement—because only together can they move us toward truly fair and trustworthy AI.

QUESTIONS?