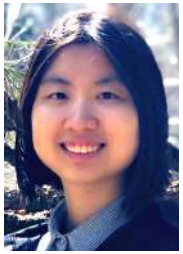
Stable Discovery of Interpretable Subgroups via Calibration (StaDISC) in Causal Studies


Raaz Dwivedi


Yan Shuo Tan


Briton Park


Mian Wei


Kevin Horgan


David Madigan


Bin Yu

Stable discovery of interpretable subgroups via calibration in causal studies.
International Statistical Review, 2020
also at arXiv:2008.10109

Effects of drugs are heterogeneous

Both in terms of efficacy and safety

Conditional Average Treatment Effect (CATE) models are used to estimate heterogeneity, but are underpowered in RCTs

Question: How can CATE models still be used to identify subgroups of patients who benefit more from the drug while having fewer side effects?

StaDISC: Applying PCS to CATE modeling

Stable Discovery of Interpretable Subgroups via Calibration

C

Feature Engineering
+ 18 CATE Models

P

Calibration-based
predictive screening

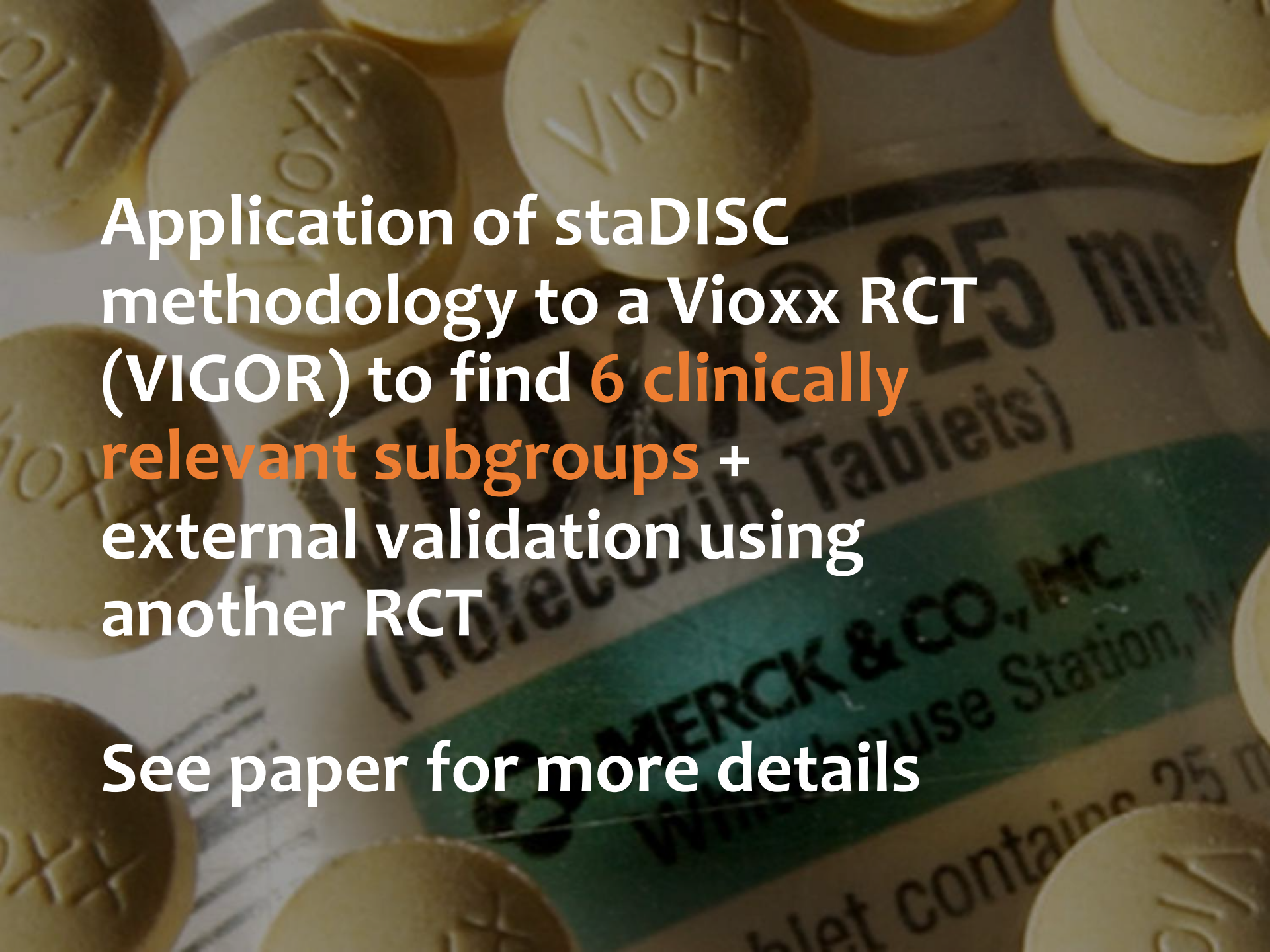
+

S

Stability to data/model/
judgment perturbations

Ranking and ensemble using
P + S checks

Finding interepretable
subgroups



Application of staDISC methodology to a Vioxx RCT (VIGOR) to find 6 clinically relevant subgroups + external validation using another RCT

See paper for more details

Data Science Book by Yu and Barter with MIT Press

Free on-line interactive copy (plan: 2022 spring)

Veridical Data Science: A Book

Bin Yu[1,2] and Rebecca Barter[1]
[1]Department of Statistics, UC Berkeley
[2]Department of Electrical Engineering and Computer Science, UC Berkeley


Berkeley
UNIVERSITY OF CALIFORNIA

What skills does the book teach?

Veridical Data Science (VDS) will teach the critical thinking, analytic, human-interaction and communication skills required to effectively formulate problems and find reliable and trustworthy solutions.
VDS explains concepts using visuals and plain English, rather than math and code.
The primary skills taught are:

Critical thinking

Readers will learn to:

Formulate answerable questions using the data available
Scrutinize all analytic decisions and results
Document all analytic decisions
Appropriate common techniques to unfamiliar situations
Deal with real, messy data

Technical skills

Data processing
Data cleaning
Exploratory Data Analysis
Data merging

Algorithmic
Dimension reduction
Clustering
Least Squares & ML
Regularization

Stability-based inference
Inference
Causal Inference
Perturbation Intervals
Trustworthiness Statements

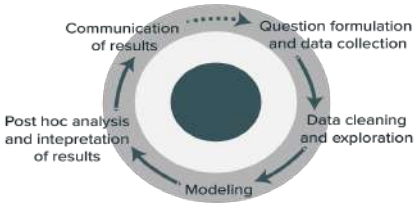
Communication

Exploratory Visual Summaries
Preparing explanatory visual and numeric summaries for explaining data and findings to an external audience

Written reports
Preparing written analytic reports for case studies based on real, messy data

Core guiding principles for the book

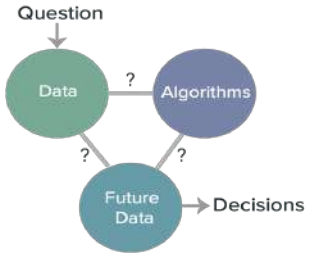
The DS Lifecycle



The Data Science Lifecycle is an iterative process that takes the analyst from problem formulation, data cleaning, exploration, algorithmic analysis, and finally to obtaining a verifiable solution that can be used for future decision-making.

Blending together concepts from statistics, computer science and domain knowledge, the data science life cycle is an iterative process that involves human analysts learning from data and refining their project-specific questions and analytic approach as they learn.

Three realms



Readers will learn to view every data problem through the lens of connecting the three realms:
(1) the question being asked and the data collected (and the reality the data represents)
(2) the algorithms used to represent the data
(3) future data on which these algorithms will be used to guide decision-making.
Guiding the reader to connect the three realms is a means of guiding the reader through the data science lifecycle.

PCS framework

The PCS framework provides concrete techniques for finding evidence for the connections between the three realms.
Predictability: if the patterns found in the original data also appear in withheld or new data, they are said to be predictable. If an analysis or algorithm finds predictable patterns, then these patterns are likely to be capturing real phenomena.
Computability: algorithmic and data efficiency and scalability is essential to ensuring that the results and solutions (e.g. a predictive algorithm) can be efficiently applied to new data.
Stability: minimum requirement for reproducibility. If results change in the presence of minor modifications of the data (e.g. via perturbations) or human analytic decisions, then there might not be a strong connection between the analysis/algorithms and the reality that underlies the data.

Intended Reader/Audience

Anyone who wants to learn the intuition and critical thinking skills to become a data scientist or work with data scientists.
Neither a mathematical nor a coding background is required.
VDS could form the basis of a semester- or multi-semester-long introductory data science university course, either as an upper-division undergraduate or early graduate-level course.

Interested? Get in touch!

Bin Yu
Email: binyu@stat.berkeley.edu
Website: https://www.stat.berkeley.edu/~binyu/Site/Welcome.html

Rebecca Barter
Email: rebeccabarter@berkeley.edu
Website: www.rebeccabarter.com
Twitter: @rlbarter