ISCTM

International Society for CNS Clinical Trials and Methodology

Biomarker identification for patient enrichment strategies in CNS clinical trials: Alternative approaches and Challenges

Dr. Joseph Geraci

NETRAMARK

drjoe@netramark.com

Disclosures

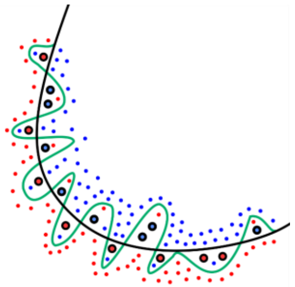
- **Netramark Corp. CSO/CTO**
- **Queen’s University, Canada**
- **Augusta University, Georgia**
- **Institute for Human Imagination, University of California**

TRADITIONAL METHODS

A Brief Review of Machine Learning methods and their limitations

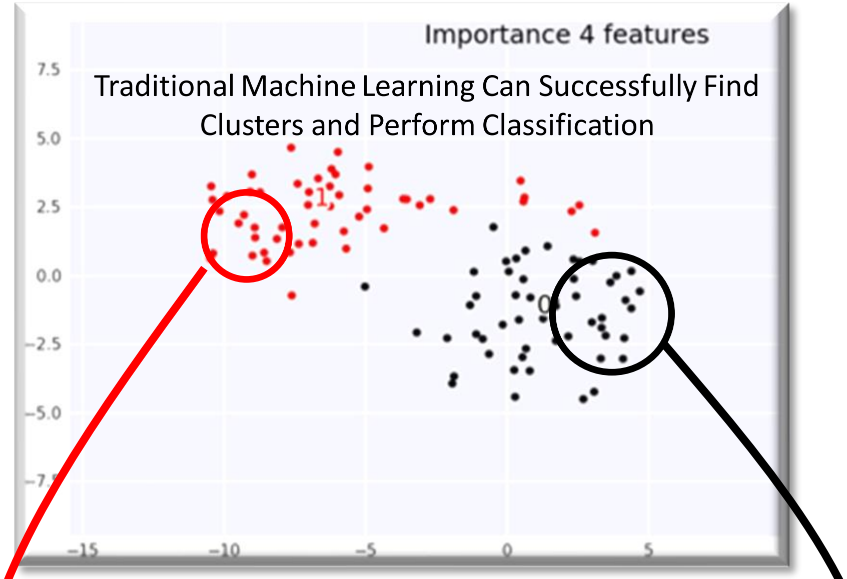
Traditional methods:

- T-Tests and ANOVA
- Chi-Square Test
- Logistic Regression
- Linear Regression
- Feature selection + Regularization
- Generalized Estimating Equations (GEE)
- Mixed Effects Models
- Random Forest
- Gradient Boosting Machines
- Support Vector Machines
- Neural Networks and Deep Learning
- Various Clustering methods like k-means, t-SNE, UMAP
- Principal Component Analysis (PCA)
- Time Series Analysis



Benefits and Risks of using traditional methods

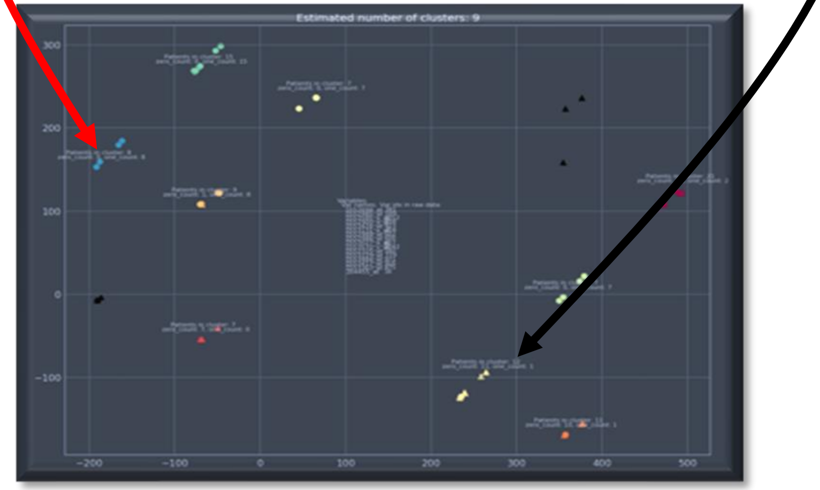
- Traditional methods are excellent when you either have a large amount of data or when objects are carefully labelled
- The problem with psychiatric patient populations is that any attempt at labelling the data and providing the machine learning methods with a dependent variable for supervised training includes ambiguity due to patient heterogeneity
- To produce insights that can eventually become biomarkers about patient populations that can generalize we need to infuse the ML methods with the ability to recognize what aspects of the data it cannot explain



Traditional Machine Learning Can Successfully Find Clusters and Perform Classification

To improve an endpoint effect size, one cannot depend on these blocky patient representations as shown above.

Which of these is closer to reality?



The heart of the challenge in understanding CNS patient populations

- Disease definitions are not precise and there are a variety of etiological manifestations that result in patient heterogeneity
- The collected variables in a clinical trial should not be expected to explain everyone.
- Response can be driven by a variety of factors
- Clinical trials provide a small number of samples which is challenging for ML, and the use of large historic databases may introduce irrelevant artifacts and drown critical nuanced factors out
- ***Machine learning methods tend to over-adhere to dependent variables***

How are methods evolving that can overcome these challenges?

- Technologies are emerging that can learn when they cannot explain certain patient subgroups, and that can decompose patient data into explainable and unexplainable parts

- The unexplainable parts provide very powerful insights about the need to collect a different modality of data

- Methods exist however that can turn insights about subpopulations into hypotheses that can be turned into enrichment criteria for clinical trials that may be at risk of failing

SUB-INSIGHT ANALYSES

Sub-Insight Analyses – Giving AI the ability to not know

- By understanding drug response and placebo response, one can use *sub-insights*, meaning highly significant hypotheses about a subset of patients, to produce biomarkers that can alter the course of a clinical trial

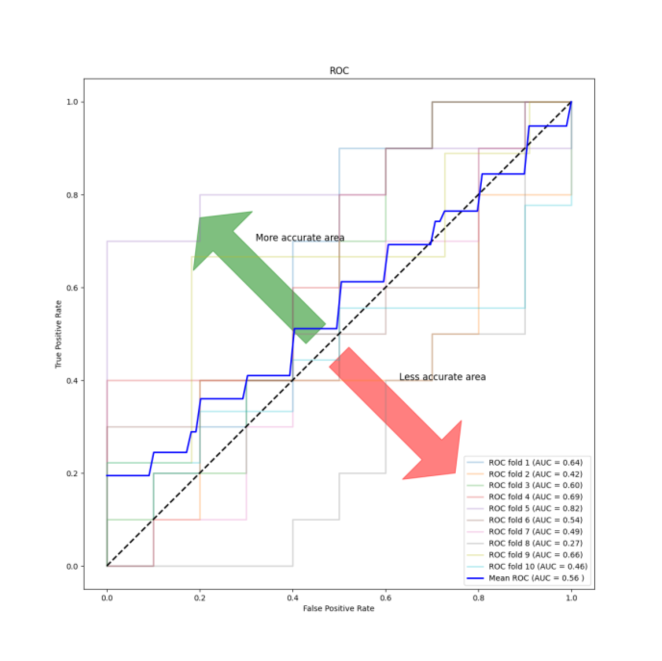
- Sub-Insights provide clinical trialists with statistically supported hypotheses about a subset of the patient population

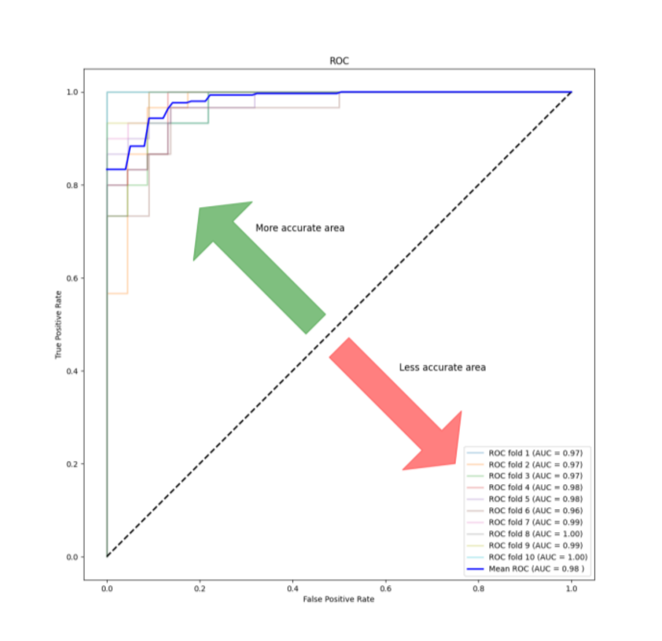
- Modelling has shown that it can be enough to utilize models that do not explain all patients but a sufficient fraction of them

- The driving variables behind these *sub-models* can then be used as exclusion/inclusion criteria

- The top models fail to replicate as can be seen by the poor set of AUCs. This is from an actual clinical trial where standard ML methods are used to try to learn about everyone in the trial wrt to drug response

- The bottom models were derived from a persona that described approximately 40% of the patients and replicated very well using standard ML methods.





From Sub-Insights to Inclusion/Exclusion Selection Criteria

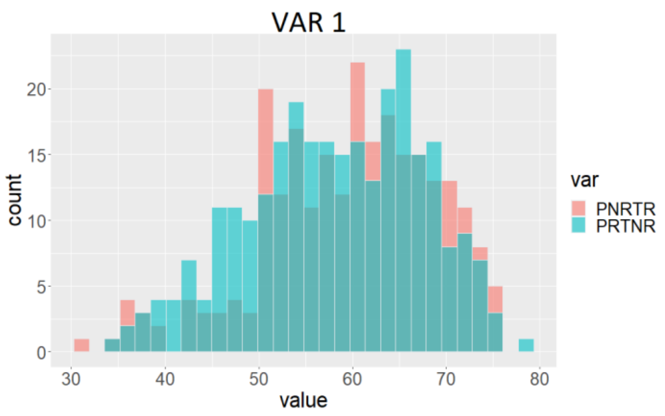
- Even though synergistic effects between learned variables are a major strength behind machine learning models, our goal here is to extract tunable parameters that trialists can use to enhance the success of their trials

- In other words, we wish to deliver transparent biomarkers, in the most general sense, to provide clear selection criteria

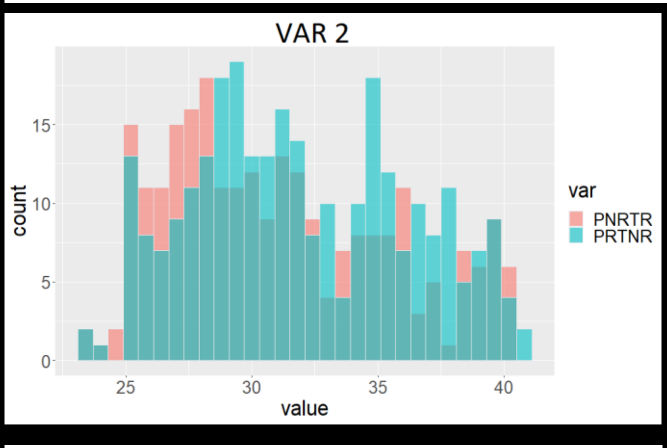
- By having the ML methods learn how to separate patients that are Placebo Responders/Treatment Non-Responders (PRTNR) vs Placebo Non-Responders/Treatment Responders (PNRTR), it can derive important factors

- These can then be studied to derive selection criteria for future trials

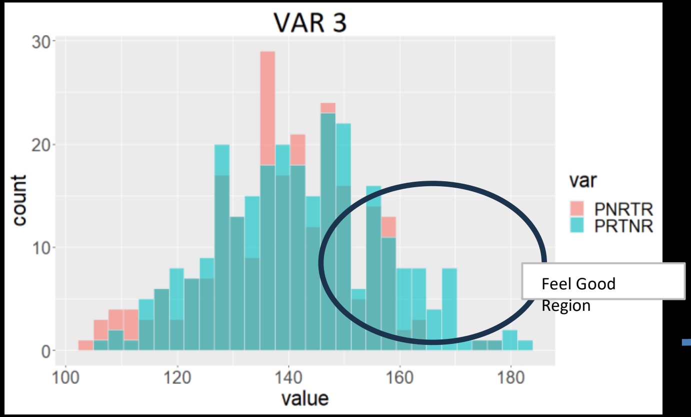
Placebo & metabolic factor



Drug Dosing and absorption factor



Placebo Response Factor



- These three variables are tunable factors that can be used to significantly improve p-values. The last variable is from a simple blood measure that physiologically corresponds to energy levels and was found to influence placebo response significantly.

Use Case Review - Phase IIa Schizophrenia Trial

- Phase II data
- Insights and patient population shattered
- Inclusion/exclusion criteria to help inform design of pivotal trials
- This use case exemplifies how this technology can be applied to Phase II trials to generate hypotheses that can inform Phase III trials enrichment criteria

USE CASE – SCHIZOPHRENIA TRIAL

Sponsor Brief:
Phase IIa Schizophrenia Trial

Data included clinical scales: CGI-S, LOF, Strauss-Carpenter Level of Functioning; mITT, PANSS, and in addition physiological measurements including heart rate, heart rate variability, positional respiration scores

138 independent variables per subject

N = 87* patients randomized into 2 arms: placebo and treatment arms with 48 in the active arm and 39 in the placebo.

Primary Endpoint: PANSS improvement (10% improvement) over placebo

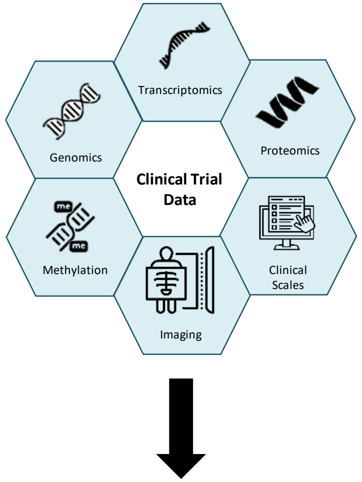
Novel medication

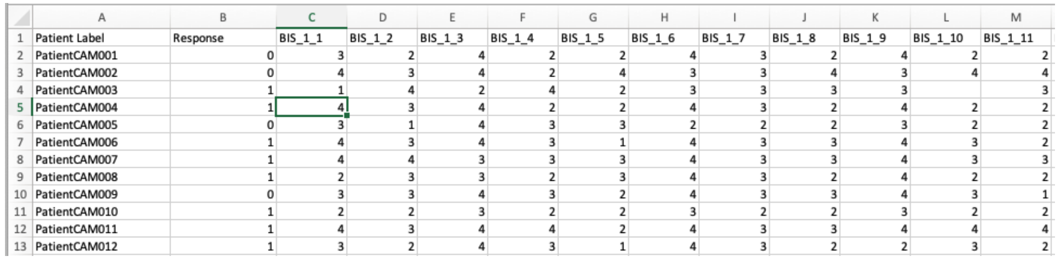
* Some patients were eliminated due to incomplete data

Project objectives

- Characterize patient response to optimize late phase design
- Mitigate risks from high placebo response which led to a marginal p-value of .04
- Establish demonstrable criteria by which to select patients before randomization to increase the certainty of demonstrating a sufficient difference of means between the placebo and active arms of the pivotal trial

Data Preparation and Ingestion





	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Patient Label	Response	BIS_1_1	BIS_1_2	BIS_1_3	BIS_1_4	BIS_1_5	BIS_1_6	BIS_1_7	BIS_1_8	BIS_1_9	BIS_1_10	BIS_1_11
2	PatientCAM001	0	3	2	4	2	2	4	3	2	4	2	2
3	PatientCAM002	0	4	3	4	2	4	3	3	4	3	4	4
4	PatientCAM003	1	1	4	2	4	2	3	3	3	3		3
5	PatientCAM004	1	4	3	4	2	2	4	3	2	4	2	2
6	PatientCAM005	0	3	1	4	3	3	2	2	2	3	2	2
7	PatientCAM006	1	4	3	4	3	1	4	3	3	4	3	2
8	PatientCAM007	1	4	4	3	3	3	4	3	3	4	3	3
9	PatientCAM008	1	2	3	3	2	3	4	3	2	4	2	2
10	PatientCAM009	0	3	3	4	3	2	4	3	3	4	3	1
11	PatientCAM010	1	2	2	3	2	2	3	2	2	3	2	2
12	PatientCAM011	1	4	3	4	4	2	4	3	3	4	4	4
13	PatientCAM012	1	3	2	4	3	1	4	3	2	2	3	2

The process

- **Data is transferred and transformed into simple tables**

- The data can include transcriptomic, epigenetics, scales, imaging, digital, real world, etc.

- **The first two columns consist of deidentified patient names and a dependent variable, i.e., the question asked. The columns to the right consist of the assorted and provided variables**

- **The number of variables for these methods can range from 20 – 1,000,000+**

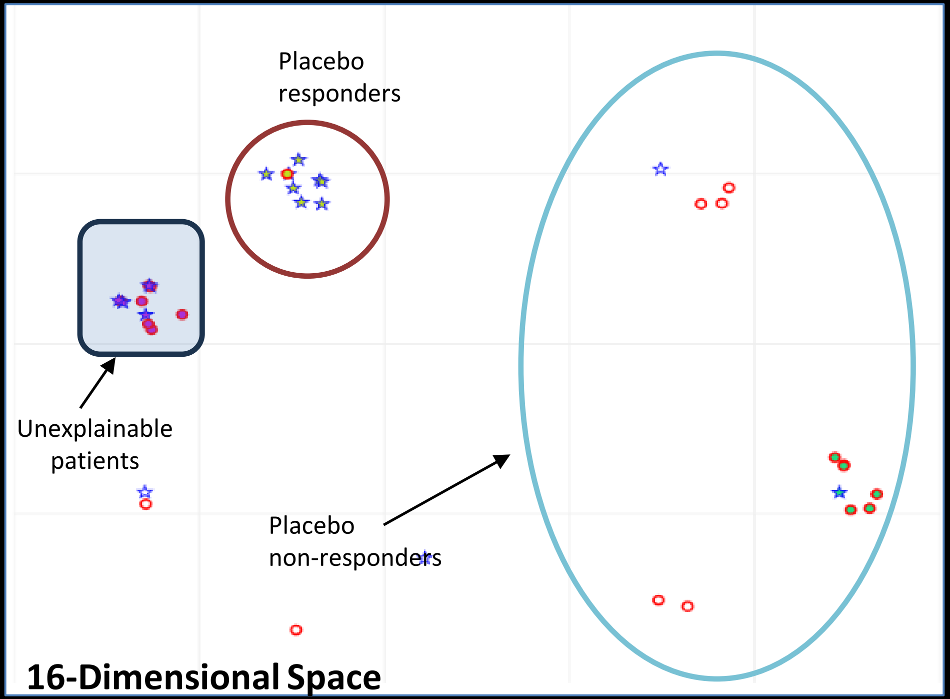
USE CASE – SCHIZOPHRENIA TRIAL

AI Analysis

The power of this approach stems from its ability to discovery subpopulations where explainable causal factors are present in combinations.

We then transform these insights into tunable parameters that can be used to increase endpoint effect size of this sponsor’s next clinical trial.

Placebo Response Hypothesis



Nearly 50% of placebo responders are characterized by having the following:

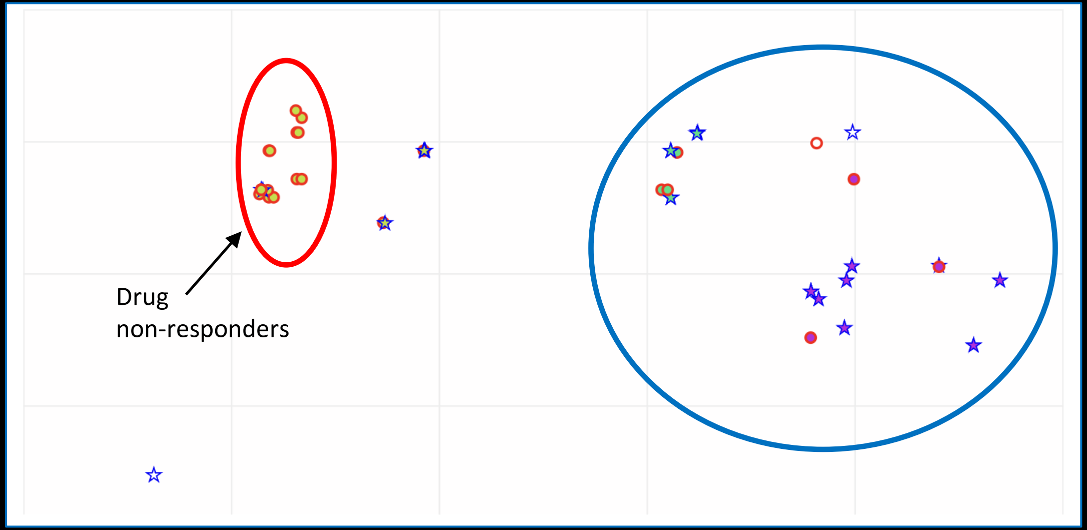
- Score less than 1 on the total baseline depression scale
- Have a supine respiration rate of 16.5 or lower
- Score 2 or higher on the emotional withdrawal item of the PANSS scale
- Score 2 or greater on the disorientation item of the PANSS scale

USE CASE – SCHIZOPHRENIA TRIAL

AI Analysis

The most powerful driving variables are discovered via a reward and elimination process. The technology uses vantages like this one to evaluate which factors are most important. Here we can see three factors that are driving drug response.

Drug Response Hypothesis

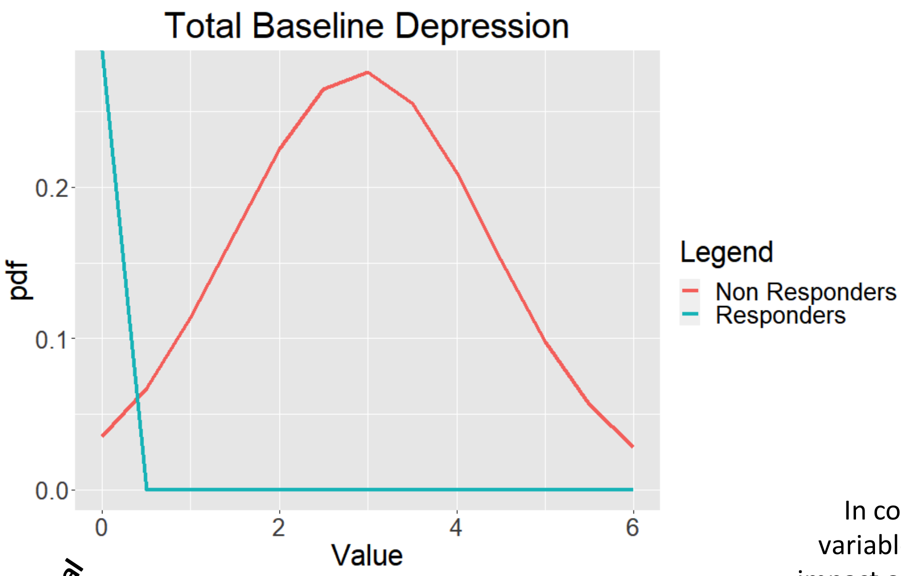


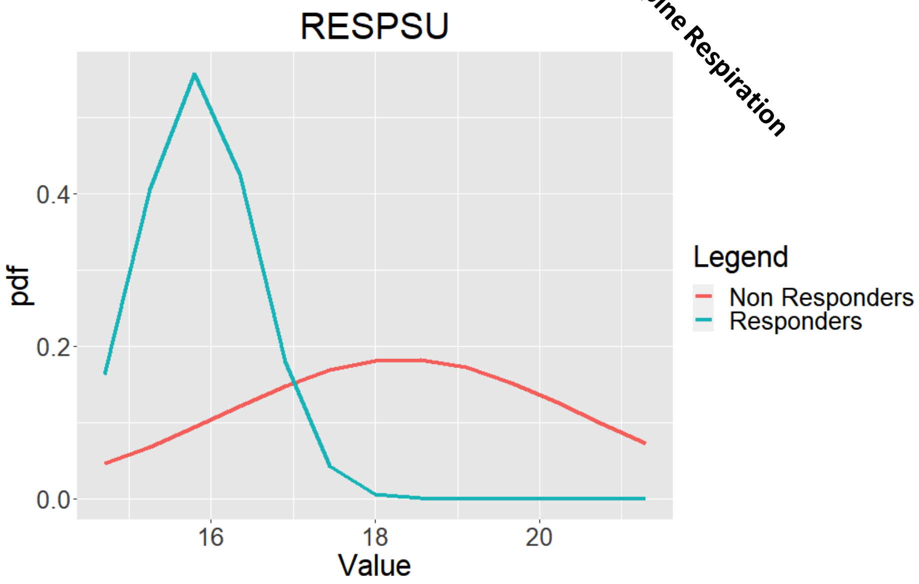
Nearly 37.5% of drug non-responders are characterized by having the following:

- A score equal to 1 (lowest) on the item corresponding to attention on the PANSS scale
- A score equal to 1 (lowest) on the item corresponding to judgement and intuition on the PANSS scale
- A score less than 5 (low) on their cognition score at baseline

USE CASE – SCHIZOPHRENIA TRIAL

Insight Delivery – Placebo Response Hypothesis

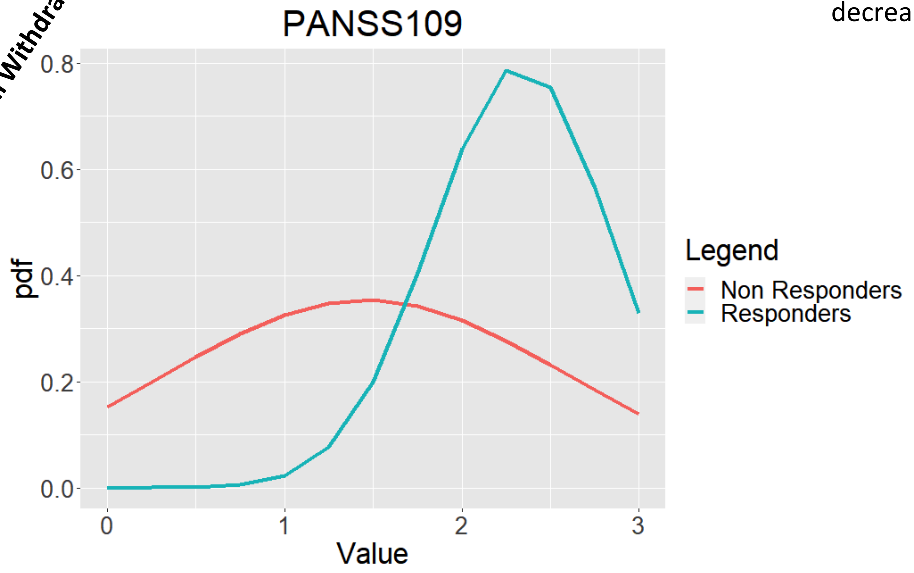

Total Baseline Depression


RESPSU

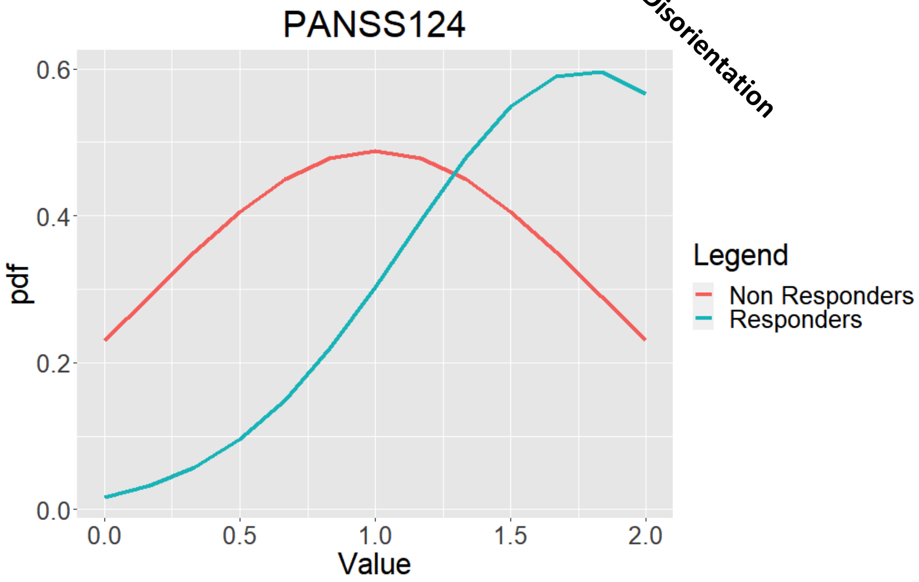
Supine Respiration

In combination, these four variables will have a significant impact on the endpoint p-value by decreasing placebo response.

Emotional Withdrawal

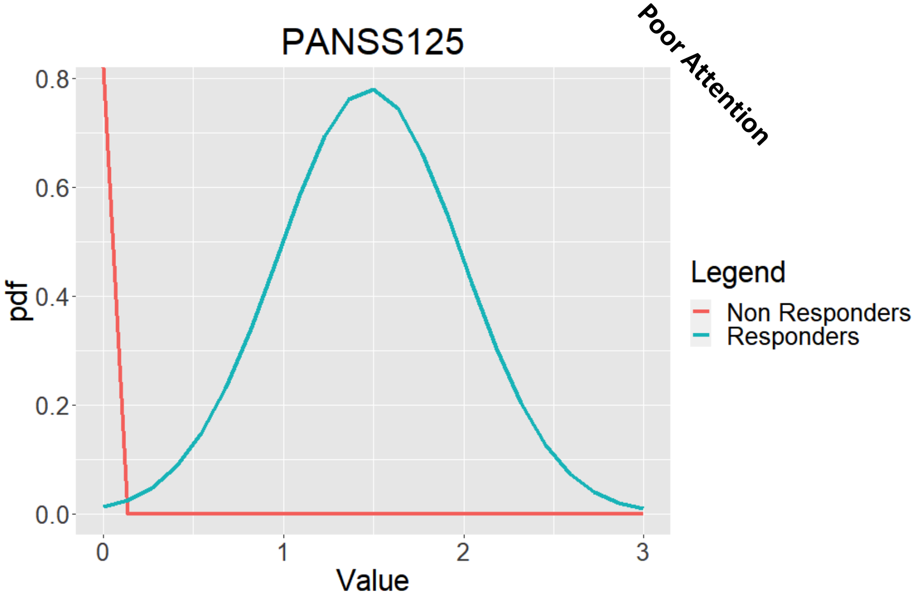

PANSS109

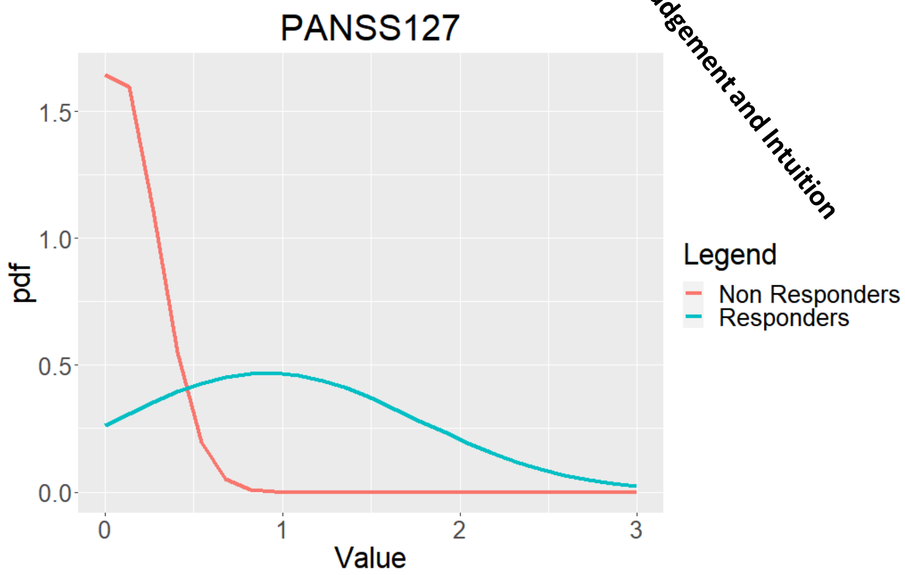
Disorientation

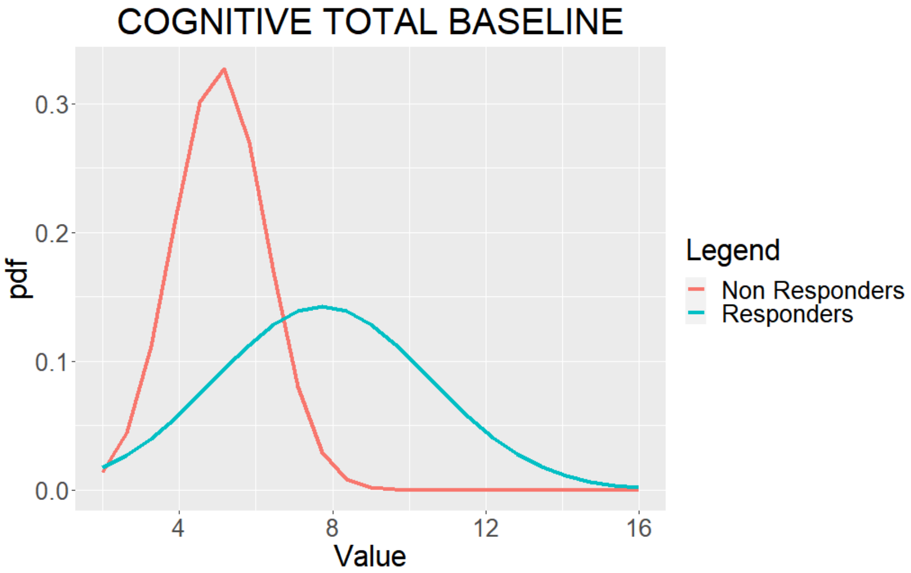

PANSS124

USE CASE – SCHIZOPHRENIA TRIAL

Insight Delivery – Drug Response Hypothesis


Poor Attention


Lack of Judgement and Intuition

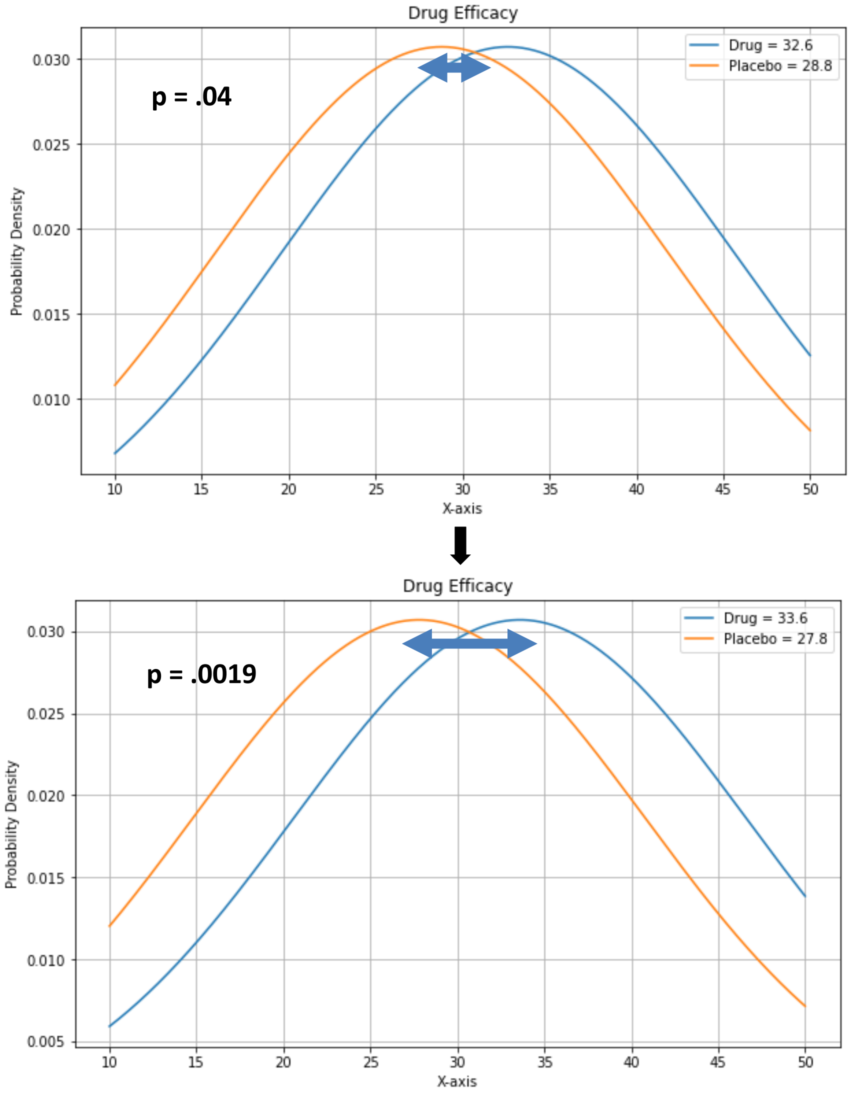


In combination, these three variables will also have a significant impact on the endpoint p-value by increasing drug response.

USE CASE – SCHIZOPHRENIA TRIAL

Even though only 30 % of the total subpopulation was explainable , the ability to take information from a highly significant set of factors allows one to alter endpoint significance

- Assumption - a scaled-up version of the clinical trial consisting of 100 people in both the active and placebo arm
- Taking the SD to be equal to 13, the difference between the placebo and active arms produced a marginally significant p value
- A mean value difference of 3.8 produced a p-value of .04
- Trial population recommendations increased the chance of success by lowering the p-value of .04 to .0019



- A significant impact on the clinical trial is possible by expecting to alter less than 20% of each arm with respect to improved efficacy in the drug arm and reduced response in the placebo arm

USE CASE – SCHIZOPHRENIA TRIAL

Potential Selection Criteria Based on NetraMark Data SCZ Study

Persona

Inclusion Criteria	Expected Effect on Study Population
Depression score ≥1 **Very Effective!**	• Removes most of the placebo responders • Little effect on drug treatment group, but removes more non-responders
Supine respiration > 16 **Questionably Effective**	• Removes most placebo responders • Removes a few more drug non-responders • High screen failure rate
PANSS Emotional Withdrawal ≤ 2 **Potentially Effective**	• Removes most of the placebo responders • Removes a more drug non-responders • Increases screen failure rate
PANSS Disorientation ≤ 1 **Effective**	• Removes many placebo responders • Removes most drug non-responders • Increases screen failure rate
PANSS Poor Attention > 0 **Effective**	• Removes drug non-responders and many PBO responders • Some increase in screen failure rate
PANSS Lack of Judgment and Intuition ≥ 1 **Effective**	• Removes most drug non-responders and PBO responders • Increases screen failure rate
Cognitive Total ≥6 **Potentially Effective**	• Removes most drug non-responders and PBO responders • Big Increase in screen failure rate

LARGE LANGUAGE MODEL INTEGRATION

Large Language Models As Insight Translators

1. **Higher Desire for Food at Baseline**

   Trial Design Insight: Consider incorporating appetite as an inclusion criterion. By selecting participants who have a higher baseline appetite, you may increase the proportion of subjects more likely to respond to the SSRI, potentially enhancing the drug’s observed efficacy in the trial.

   Measurement: Use validated appetite assessment scales or questionnaires at the screening phase.

2. **Satisfaction with Leisure Activities at Baseline**

   Trial Design Insight: Evaluating the baseline satisfaction with leisure activities can help in stratifying participants based on their potential to respond. Including individuals who still derive pleasure from activities might lead to a higher overall response rate in the trial.

   Measurement: Use psychometrically sound scales assessing anhedonia or leisure activity satisfaction during participant screening.

3. **Satisfaction with Mood at Baseline**

   Trial Design Insight: It may seem counterintuitive to include patients with some degree of mood satisfaction in a MDD trial, these individuals may represent a segment that responds particularly well to SSRIs. Stratify participants based on their mood satisfaction scores to identify differential drug responses.

   Measurement: Implement standardized mood assessment tools at baseline, ensuring the tool captures nuances in mood satisfaction.

4. **Less Enjoyment from Family at Baseline**

   Trial Design Insight: Participants with significant familial or interpersonal stressors might represent a group where SSRIs demonstrate a pronounced effect, possibly due to the drug’s buffering effect against these stressors. Consider creating a stratification analysis for participants with familial stressors or dissatisfaction.

   Measurement: Employ interpersonal relationship scales or family-related quality of life assessments during the screening phase.

Integrated Trial Strategy:

To enrich your clinical trial, utilize these predictors as stratification or subgrouping factors. This approach can help in identifying specific segments of the depressed population where the SSRI demonstrates maximum efficacy. Furthermore, these predictors can aid in patient selection, ensuring a higher likelihood of observing positive treatment outcomes, and consequently enhancing the power and validity of the trial results. Additionally, understanding these factors upfront can assist in post-hoc analyses and interpretations, helping to delineate why certain participants responded better and informing future trial designs or post-market strategies.

Challenges

- A real challenge is that academic research suggests that the best way to utilize the combinatorial benefits of machine intelligence-based models for pre-randomization patient enrichment is to apply ML derived models to this task. Regulators however require an intense evaluation process to implement models even though the risk is on the sponsor if implemented pre-randomization.

- To combat this, practitioners derive univariate tunable parameters for enrichment purposes. The main problem with this is that some of the emergent power that comes from multi-dimensional machine learning models is lost.

- State of the art deep neural networks can encode models into a complex substrate that is difficult to interpret. Cutting edge efforts reported in this presentation are designed ***to not depend on deep neural networks***, as they will also overfit due to the limited number of samples provided in clinical trial data sets, and further, explainability is a primary concern for this use case.

- The main challenge with these techniques designed for use directly from Phase II or III data, is that the insights may reflect artifacts directly within the data. Due to the clear audit path and explainability, these risks can be mitigated, and essentially the decision must be made by the clinical trialists who consider the generated hypotheses.

- Inclusion/Exclusion criteria derived from machine intelligence algorithms may discover tunable factors that actually reflect real phenomenon found among patients, but they may not be feasible to implement as they make screening too restrictive.

- The use of Large Language models runs the risk of producing fictional conclusion. However, this step does not analyze the data but only interprets what the previous steps discover, so these results are auditable if implemented correctly.