Application of machine learning algorithms to identify subjects at risk of within PANSS logical errors

Submitter Alan Kott

Affiliation Signant Health

SUBMISSION DETAILS

I agree to provide poster pdf for attendee download. Yes

Poster PDF for download <blank>

What is the Methodological Question Being Addressed? Can application of machine learning algorithms using pre-randomization subject data identify subjects at high risk of within PANSS logical errors after subject is randomized?

Introduction Data quality monitoring programs identify with various degrees of accuracy and clinical relevance quality concerns in the data as these are collected. By definition these programs are reactive as any action can be taken only after the error is identified. Implementing these programs coupled with other solutions, such as intelligent design in eCOA development and independent review of audio recordings of PANSS interviews followed by targeted remediation has proven to effectively reduce the amount of data concerns. (Kott, Brannan et al., 2020) Despite these measures a non-negligible proportion of data remains affected by various types of data concerns. Within PANSS logical errors represent a group of high frequency data concerns that signal inappropriate use of the scale and are associated with both bias and noise in clinical trial data. For example, within PANSS logical errors have been previously shown to increase response to placebo in acute schizophrenia clinical trials. (Kott, Lee et al, 2016) Machine learning (ML) offers the possibility to proactively identify future data quality concerns. In the current analysis we explore the application of ML to identify subjects at risk of within PANSS errors after the subjects get randomized into the trial using only pre-randomization data.

Methods Data were retrieved from 4,006 subjects who participated in 17 acute schizophrenia clinical trials. Data were randomly split into a training dataset (75% of all data) and test dataset (25% of the data). The model was trained using only screening PANSS and CGI-S data and a “hit” was labeled if there was at least one post-baseline within PANSS error identified. The performance of the algorithm was assessed using confusion matrix, and area under the curve from obtained Receiver operation curve. R Superlearner algorithm was utilized to train the model.

Results At least one within PANSS error was identified in 4,169 (9.9%) visits. 758 (18.8%) subjects in the training set were affected by post-baseline within PANSS errors and 263 (26.2%)subjects in the testing set. The AUC value was estimated to 0.82 indicating an excellent performance of the model. Confusion matrix was obtained for a set of predefined specificity cut-offs (70%, 80%, 90%, 95% and 99%). For example – for the specificity set at 90% the model correctly identified 237 hits (true positives) and incorrectly 374 hits (false positives), and it missed 26 hits (false negatives). The true negatives represented 365 subjects.

Conclusion Our results suggest that ML can with excellent performance predict post-baseline within PANSS logical discrepancies using only screening data. The model was trained using only individual PANSS item scores and CGI-S score thus offering a general applicability across acute schizophrenia clinical trials. Once risks on an individual subject level are identified, it is critical to identify raters and sites where these risks are accumulated with outlying frequency and subsequently perform a targeted remediation.

Co-Authors

* Presenting Author

First Name	Last Name	Affiliation
Alan *	Kott *	Signant Health
Xingmei	Wang	Signant Health
David G.	Daniel	Signant Health

Keywords

Keywords
Machine learning
PANSS
Data quality

Guidelines I have read and understand the Poster Guidelines

Disclosures All authors are full time employees of Signant Health. The poster was financially supported by Signant Health.

Related Tables and Supporting Materials <blank>