Semi-automated Screening for Improbable Randomization in PDFs

Colby J. Vorland, David B. Allison, Andrew W. Brown

Indiana University School of Public Health-Bloomington

Keywords: randomization, automation, PDFs, table extraction, software

Background

The use of statistical principles has been applied to uncover cases of large-scale fraud or misreporting of methods of published randomized controlled trials (RCTs). If the distribution of p-values for group comparisons of baseline characteristics is skewed (i.e., not uniformly distributed) or if there are very small p-values in a group of papers by one author group, it may indicate that randomization was implemented incorrectly or not at all, for which methodological inquiries to the authors may be warranted. For instance, Carlisle applied this screening approach to uncover 168 fraudulent RCTs by Yoshitaka Fujii [1]. Bolland and colleagues flagged 33 RCTs of Yoshihiro Sato [2], who admitted to faking data, and recently flagged 172 papers by a nutrition researcher that appear suspect [3], many of which now have associated expressions of concern and are being investigated. Our group has also used this method to point out anomalies, leading to retractions [4,5]. Thus, the method is useful to provide statistical evidence to inform research integrity investigations of published RCTs.

Contribution

Manually screening many RCTs and extracting baseline p-values, or manually calculating p-values from summary statistics if p-values are not presented, is burdensome. Automated methods may be able to assist those screening papers for improbable randomization, and a preliminary evaluation of such a process is presented herein.

Data and Methods

To upload PDFs for processing, display results and interact with the PDFs, a Flask-based web application was developed. The application loads the PDF while extracting information in the background, and outputs results in a panel to the left of the PDF (**Figure 1**).

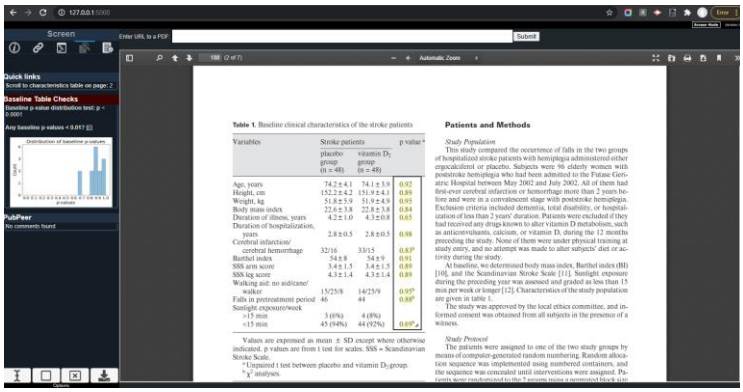

Figure 1. Software interface.

Once a PDF is uploaded, pages are converted to images and a convolutional neural network algorithm trained to detect tables is applied to each image to identify potential table regions. Coordinates are then passed to a table extraction software package to extract the structured tables as data frames, and custom rule-based Python functions look for indications that the table contains baseline characteristics. If a baseline table is identified, a function attempts to extract the p-value column if one exists. If one is not found, additional rule-based functions look for group numbers in the table header and error type (i.e., standard error or standard deviation), and if found, computes p-values for each row using a t-test or Fisher exact test for continuous or categorical variables, respectively for trials with 2 groups. The Stouffer-Fisher method is used to calculate an overall p-value of the

uniformity of the distribution, and the p-values are plotted graphically; both results are shown to the user. Drag/drop functions over the PDF allows the user to manually highlight columns or rows of information to perform calculations if all required information cannot be identified automatically.

For a preliminary evaluation of this process, the set of 33 RCTs by Yoshihiro Sato and colleagues previously identified as suspect using the baseline p-value method [2] was used to explore whether an automated approach would flag these trials.

Results

Baseline tables were successfully identified and extracted in 26/31 PDFs (84%) (**Figure 2**). Of these 26, 15 trials (48%) were successfully assessed in a fully automated manner; 13 were flagged as improbable by a low (=< 0.02) p-value from the Stouffer-Fisher test or by the presence of low p-values (< 0.01) published in the baseline table. Of the 15 trials screened automatically, 12 reported all p-values needed to perform assessments, and 3 had none reported or a mix of exact and non-exact p-values (i.e., "NS" or "<0.05") for which summary statistics were used to automatically compute them.

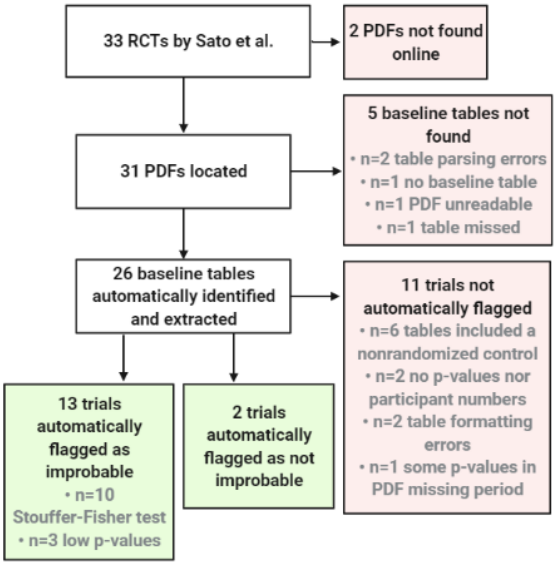

Figure 2. Results of automated screening of the RCTs by Sato et al.

Of the 11 trials not flagged, their evaluation could all be expedited with the combination of drag/drop functionality and manual inputs where needed (e.g., if participant numbers are not listed in the table).

Impact

Based on these preliminary results, this semi-automated process may save considerable time when screening a large group of papers for improbable randomization by automating as much as possible and displaying the results to an investigator for verification. In addition, it may have utility for general surveillance of the broader literature, with human follow-up to interrogate methods of papers that are flagged. Additional work will also evaluate automated screening of tables for other indicators of data fabrication, such as digit preference.

References

1. Carlisle J. The analysis of 168 randomised controlled trials to test data integrity. *Anaesthesia.* 2012;67(5):521-537.
2. Bolland MJ, Avenell A, Gamble GD, Grey A. Systematic review and statistical analysis of the integrity of 33 randomized controlled trials. *Neurology.* 2016;87(23):2391-2402.
3. Bolland MJ, Gamble GD, Avenell A, Grey A. Rounding, but not randomization method, non-normality, or correlation, affected baseline P-value distributions in randomized trials. *Journal of clinical epidemiology.* 2019;110:50-62.
4. Mestre LM, Dickinson SL, Golzarri-Arroyo L, Brown AW, Allison DB. Data anomalies and apparent reporting errors in 'Randomized controlled trial testing weight loss and abdominal obesity outcomes of moxibustion'. *BioMedical Engineering OnLine.* 2020;19(1):1-3.
5. George BJ, Brown AW, Allison DB. Errors in statistical analysis and questionable randomization lead to unreliable conclusions. *J Paramed Sci.* 2015;6(3):153-154.