Integrity Analysis: An open-source program to institute the Carlisle-Shafer method of analyzing baseline data from a randomized controlled trial.

Background:

For about 15 years papers published by Yoshitaka Fujii had been considered sketchy. In 2000 Kranke and colleagues wrote a snarky letter to Ronald Miller, Editor-in-Chief of Anesthesia & Analgesia, with the sarcastic title: “Reported data on granisetron and postoperative nausea and vomiting by Fujii et al. Are incredibly nice!” The authors pointed out the impossibility of nearly every group in every randomized controlled trial published by Fujii having one headache as an AE. Dr. Fujii’s response was “the data are the data”, and nothing more was done.

In 2012 John Carlisle published a landmark paper “The analysis of 168 randomised controlled trials to test data integrity”.[[1]](#endnote-1) Carlisle’s insight was that the baseline data of a randomized controlled trial are all samples of the same pre-treatment population. The difference between samples is a function of the raw data, and can be inferred from the standard deviations of the measurements. One can test whether the means of two groups are “too close”.

For example, let’s say the mean weight of the control group of 6 subjects is 77 ± 30 kg and the mean weight of the treatment group of 6 subjects is 78 ± 30 kg. If the standard deviation is 30, then the standard error of the mean is ~8 (30/√6). The mean of each group is just 0.5 kg from the mean of 77.5. A two-tailed T test yields a p value of 0.96. That says that 96% of the time one would expect the means to be this far apart, or further apart. OK, they aren’t far enough apart to indicate that they are from different populations. However, are they too close? Well, if 96% of the time they are this far apart or further apart, then only 4% of the time will they be this close, or closer. So there is only a 4% chance that this is just a result of random sampling.

If you see this once, it’s no big deal. It’s expected in 4% of random samples! However, if you see this again and again, then something is definitely fishy with the data.

In reviewing 168 papers by Yoshitaka Fujii, Carlisle found that the mean of the baseline values were very often too close, based on the standard error of the mean for each row in the baseline tables. The P value for all results approached 1 in the number of atoms in the universe.

Carlisle’s insights and analyses unmasked years of data fabrication by Fujii and retraction of 172 papers at last count.[[2]](#endnote-2) However, there were two problems. First, the conventional definition of standard deviation is the square root of the variance. Unfortunately, that isn’t quite right. Although the definition of variance is unbiased, taking the square root of the variance produces a biased estimate of standard deviation. The bias is modest,[[3]](#endnote-3) and using a biased estimate of standard deviation did not affect John Carlisle’s results.

A more significant problem is introduced by rounding. Let’s say that in our prior example both groups had a mean weight of 77. Unlikely, but possible. However, the analysis falls apart. The difference between the groups is 0. As a result, the P value for that is 1 (all possible differences in mean must be 0 or greater). The P value for a difference of 0 or less is 0. In other words, the likelihood is infinitely small. That makes no sense in the real world, because data are rounded. However, from the perspective of normal statistical theory, a difference of 0 between two random samples from a population is impossible.

The only way I know to handle this is to completely reject normal statistical theory, and instead replicate the study using Monte Carlo simulations of the study including the requisite rounding of data. John Carlisle and I spent several years developing this method. In 2015 we published a re-analysis of the Fujii data.[[4]](#endnote-4) The results were the same. However, in simulation we demonstrated that the Monte Carlo approach was more robust than using normal theory.

The basic idea of the Monte Carlo technique is that the baseline table in an RCT is entered as a spreadsheet. Each row of the spreadsheet corresponds to a single row / column entry. For example, in our weight example, we might have



This says that there are two entries for weight. One is 77 ± 30, and the other is 78 ± 30. Both are from populations of just 6 subjects. The weight isn’t rounded (it’s just an integer), but the MEAN is rounded to 1 decimal point.

One can then simulate a study based on the table thousands of times. Once the simulations are done, one can then ask how often are the weights just 1 kg or less different from each other? If that is a rare event (say just 4% of the time), then the numbers are “too close” and potentially sketchy.

This is the basis of the Shiny program “IntegrityAnalysis” that I have made available online (see <https://steveshafer.shinyapps.io/IntegrityAnalysis/>).

**How to use the Integrity Analysis program**

1. Prepare your spreadsheet. Excel (xlsx and xls) and csv file formats are supported.
2. Upload your spreadsheet into the program
3. Analysis starts as soon as the spreadsheet is uploaded. By default 100,000 bootstrap replications are performed for each group in the Monte Carlo analysis.
4. Download the results.
5. The results file includes the original file, with an additional column “P”. The P value is shown for each row.
6. The “Overall P” value for the trial appears at the bottom. It is computed for using the sumz() implementation of Stouffer’s method of combining p values.[[5]](#endnote-5)

**How to prepare your data file: continuous variables**

The column named “ROW” describes what is being measured (e.g., weight, height, age, sex, surgery, etc). It is required for every line to identify what, exactly, is reported on the line.

Three entries are required for every continuous variable: MEAN, SD, and N. MEAN is the algebraic mean of the data, as would typically be found in a table of baseline values. SD is the standard deviation of the data (e.g., the square root of the variance). N is the number of subjects in the group.

An additional entry, ROW, is the label of the row in the table of baseline values. It is alphanumeric (e.g., “Weight”, “Height”, “Age”). If there are three groups in a study, then the each baseline variable should be listed 3 times, one in each row.

Returning to our previous table, if the study had 3 groups, then the baseline values of weight might look like this:



This means that there are three experimental arms. Each has 6 subjects. It’s a bit sketchy to have the N and SD identical in all three experimental arms, so perhaps a more realistic example is:



This might suggest that there are two treatment arms of 50 subjects each, and a control arm of 100 subjects.

IntegrityAnalysis can analyze multiple trials in a single run. If there is no field called “TRIAL” in the spreadsheet, then IntegrityAnalysis will add a column called “TRIAL”, and assign it the number 1. If there are several trials in the spreadsheet they must be identified in a column labeled “TRIAL”. For example, let’s say we have two trials, A and B. Trial A has 3 groups. Only weight is reported in the baseline data table. Trial B has 2 groups and only reports age. Here is how that spreadsheet might appear.



Trials are analyzed independently. For example, the analysis for trial A above will have no effect on the analysis for trial B (and vice versa, of course).

A row in a baseline table is defined by the trial name and the row name. In the above example, there are 3 rows for Trial “A”, Row “Weight”. These will be analyzed together. They do not need to follow each other in the spreadsheet. For example, this spreadsheet will return the same result as the above spreadsheet:



The name of the row can appear for multiple trials. For example, most trials will report weight and age for each group in the baseline table. The weight of subjects in trial A will have no influence on the weight of subjects in trial B.

Continuous data must have a MEAN, N, and SD defined for each item. If any of these are omitted, the program will identify the error and ask the user to upload a corrected spreadsheet.

There are two fields for rounding: Round Observations and Round Mean. If these are omitted, the program will attempt to add these based on the decimal places for the mean values. In the Monte Carlo simulations, the simulated observations are rounded using the ROUND OBSERVATION” entry. The mean of these is taken, and rounded to the “ROUND MEAN entry.”

**How to prepare your data file: categorical variables**

Categorical variables are count variables, often used to include sex, race, and ethnicity. These appear as additional columns in the spreadsheet. This is from the example spreadsheet.



There appear to be two groups in the study. The first group has 40 men and 10 women. The second group has 34 men and 16 women. Evidently there were three kinds of surgery: upper, lower, and urologic.

When categorical variables are present, MEAN, SD, and N should be blank. If they are not blank, the program will point out the error and ask that it be corrected.

The P values for categorical variables are also determined using Monte Carlo simulation. Fortunately, the chisq.test() function in R specifically allows the use of Monte Carlo simulation to determine the P values of categorical variables using “**simulate.p.value = FALSE, B = 100000”**.

**Spreadsheet processing**

IntegrityAnalysis does a moderate amount of processing to catch errors in the spreadsheet and make thoughtful guesses for ambiguous entries. Specifically:

1. All column names are converted to uppercase characters, and leading / trailing whitespace is removed.
2. The first column containing the text “TRIAL” is changed to “TRIAL”.
3. The first column containing the text “ROW” is changed to ROW.
4. The first column containing the text “MEAN” is changed to “MEAN”. If there is an additional column containing the text “MEAN”, it is changed to “ROUND MEAN”.
5. Columns labeled N, MEAN, and SD must be present, even if the baseline data comprises categorical variables only.
6. The first column containing “OBS” is changed to “ROUND OBSERVATION”
7. Columns containing categories are identified by:
   1. Having at least one entry
   2. Not having an entry for every line (which would be unusual for a study)
   3. Comprising only integer values
8. Columns other than TRIAL, ROW, N, MEAN, SD, ROUND MEAN, ROUND OBSERVATION, and identified categories are ignored.
9. Columns with non-blank values of N, MEAN, and SD cannot contain entries for categorical variables.
10. Columns with non-blank values of any categorical variable cannot contain entries for N, MEAN, and SD.
11. In 2017 John Carlile published an analysis of 5087 randomized controlled trials.[[6]](#endnote-6) John generously provided the spreadsheet of continuous variables used in his analysis. To accommodate his column names:
    1. Any column labeled “MEASURE” is renamed “ROW”, and the column labeled “GROUP” is deleted.
    2. Any column labeled “DECM” is renamed “ROUND MEAN”.
    3. Any column labeled “NUMBER” is renamed “N”.
    4. His Carlisle JB. Data fabrication and other reasons for non-random sampling

**Processing**

File validation starts as soon as the file is uploaded. An error message will appear if any issues are identified.

If no issues are identified, IntegrityAnalysis will immediately begin the Monte Carlo simulations. These may take several minutes to complete. Unfortunately, it is exceptionally difficult to output a message to the user with the status of the analysis.

1. # in 5087 randomised, controlled trials in anaesthetic and general medical #
2. # journals. Anaesthesia. 2017;72:944-952

Similarly, the Row name can

TRIAL ROW MEAN N SD ROUND OBSERVATIONS ROUND MEAN  
A Weight 77 6 30 0 1  
A Weight 78 6 30 0 1  
A Weight 64 6 30 0 1

A Weight 77 6 30 0 1  
A Weight 78 6 30 0 1  
A Weight 64 6 30 0 1

, then the differe kg 32, and the standard deviation is

If the reported means of the data are “too-close”,

(Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia. 2012;67:521-537.)

1. Carlisle JB. The analysis of 168 randomised controlled trials to test data integrity. Anaesthesia. 2012;67:521-537. [↑](#endnote-ref-1)
2. https://retractionwatch.com/the-retraction-watch-leaderboard/ [↑](#endnote-ref-2)
3. https://en.wikipedia.org/wiki/Unbiased\_estimation\_of\_standard\_deviation [↑](#endnote-ref-3)
4. Carlisle JB, Dexter F, Pandit JJ, Shafer SL, Yentis SM. Calculating the probability of random sampling for continuous variables in submitted or published randomised controlled trials. Anaesthesia. 2015;70:848-58 [↑](#endnote-ref-4)
5. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RMJ (1949). The American soldier, vol 1: Adjustment during army life. Princeton University Press, Princeton. [↑](#endnote-ref-5)
6. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia. 2017;72:944-952 [↑](#endnote-ref-6)