Algorithm validation: Comparison of the statistics extracted from baseline tables by the algorithm and manually

Adrian Barnett

08 June 2022

This document examines the accuracy of the automated algorithm for extracting table data by comparing it with manually entered data for a random selection of 200 trials.

# Exclusions

In this section we look at the reasons tables were excluded.

### Excluded by algorithm

| **algorithm** | **n** | **percent** |
| --- | --- | --- |
| Full text page not available | 51 | 40 |
| No baseline table | 39 | 31 |
| Just one column in table | 18 | 14 |
| No sample size | 6 | 5 |
| Follow-up results in baseline table | 5 | 4 |
| Just one sample size | 3 | 2 |
| Single-arm study | 3 | 2 |
| Pre-post comparison | 1 | 1 |
| Total | 126 | - |

### Excluded by manual entry

| **manual** | **n** | **percent** |
| --- | --- | --- |
| No baseline table | 38 | 44 |
| Just one column in table | 26 | 30 |
| Full text page not available | 19 | 22 |
| Study protocol | 2 | 2 |
| Follow-up results in baseline table | 1 | 1 |
| Graphical table | 1 | 1 |
| Total | 87 | - |

More trials were unavailable to the algorithm than the hand-entered data as the XML file was not available whilst the online HTML version was accessible. This could be investigated further to increase the number of trials available to the algorithm.

Some trials were not listed as excluded in the manually entered results, e.g., those with graphical tables (PMC7310966). The algorithm can not process graphical tables, whereas a human reader can.

### Cross-tabulation of exclusion reasons

The cross-tabulation below is the exclusions for the manually entered data (rows) and algorithm (columns). The grey cells show where the exclusions reasons are the same for both methods.

| **manual** | **FURIBT** | **FTPNA** | **JOCIT** | **JOSS** | **NBT** | **NSS** | **PPC** | **Not excluded** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Follow-up results in baseline table | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Full text page not available | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 1 |
| Graphical table | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Just one column in table | 0 | 3 | 14 | 2 | 6 | 1 | 0 | 0 |
| No baseline table | 0 | 10 | 2 | 1 | 24 | 0 | 1 | 0 |
| Study protocol | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| Not excluded | 4 | 20 | 5 | 0 | 6 | 5 | 0 | 0 |

Acronyms explained:

* FTPNA = Full text page not available
* FURIBT = Follow-up results in baseline table
* JOCIT = Just one column in table
* NSS = No sample size
* NBT = No baseline table
* PPC = Pre-post comparison
* JOSS = Just one sample size

As mentioned above, there were a relatively high number of trials that were available online for manual entry but not in XML format for the algorithm (green cell).

The algorithm failed to detect six baseline tables (yellow cell). Two tables were in online supplements which the algorithm did not search (PMC6964911, PMC7919220), two papers had no baseline table (PMC7698004, PMC7400245), one paper had a graphical table that the algorithm cannot process (PMC7310966), and one paper had a table legend that did not clearly identify it as a baseline table (PMC7468014).

There were five studies where the manual entry managed to extract the sample size but the algorithm did not (red cell). This was because the sample size was not in the table, title or footnote.

# Table data

This section examines the differences in the table data extracted by the algorithm compared with the manually entered data.

## Differences in row and column numbers

| **Number of tables** | **Median number of rows (manual)** | **Median number of columns (manual)** | **Median difference and 90% interval for rows** | **Median difference and 90% interval for columns** |
| --- | --- | --- | --- | --- |
| 74 | 14 | 2 | 0 (-1, 6) | 0 (0, 0) |

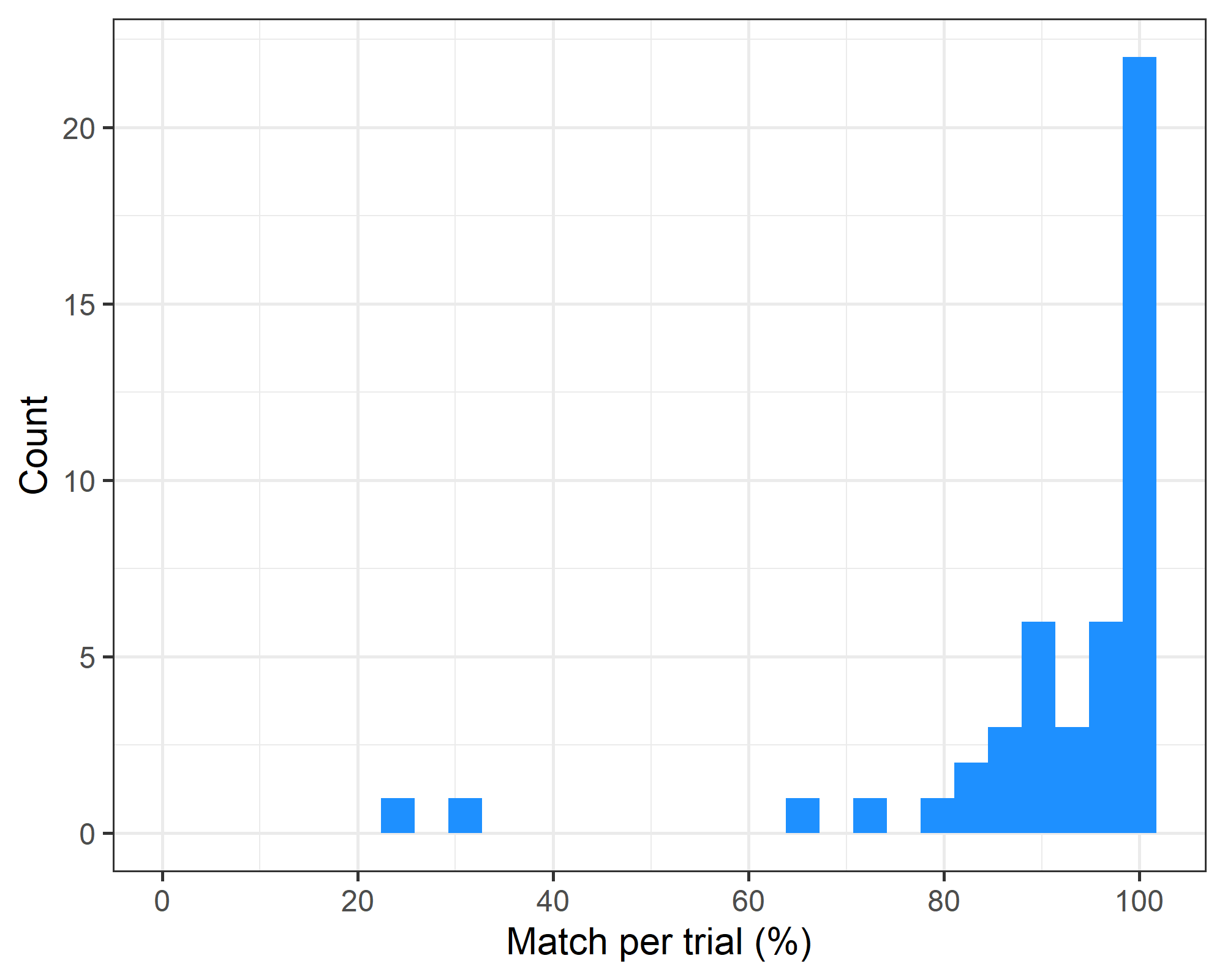
The aim of this table is to examine differences in the sizes of the tables.

The difference is the manual-entered data minus the algorithm. The sizes of the tables were comparable. The manual-entered data sometimes had more rows that the algorithm. This is because the algorithm excluded rows where the written percentages did not agree with a re-calculation using the numerator and sample size.

## Differences in statistics type

This section looks at the estimated statistics type, such as “continuous” or “percent”. I combined “percents” and “numbers” as are treated the same in the analysis. I excluded 71 results where there was no statistic given for the manually entered data.

### a) Differences per trial



Poor matching was sometimes due to:

* The algorithm rejected multiple rows in the table data as the written percentages had not been calculated correctly (PMC7038258, PMC6966838, PMC6986109).
* The table had badly aligned columns (PMC7424488).
* The table was split into multiple sub-tables and the algorithm only captured the first table, creating missing data (PMC7528007).

### b) Differences by statistic type

|  | **Algorithm** | | | | |
| --- | --- | --- | --- | --- | --- |
| **manual** | **confidence interval** | **continuous** | **median** | **percent** | **Missing** |
| confidence interval | 0 | 1 | 2 | 4 | 1 |
| continuous | 0 | 206 | 2 | 3 | 13 |
| median | 2 | 0 | 55 | 0 | 3 |
| percent | 0 | 12 | 0 | 482 | 42 |
| Missing | 0 | 3 | 0 | 1 | 0 |

A large number of mismatches come from percents that the algorithm classified as “missing”, likely because the numerator divided by the sample size did not give the percent, and hence the algorithm deleted the number.

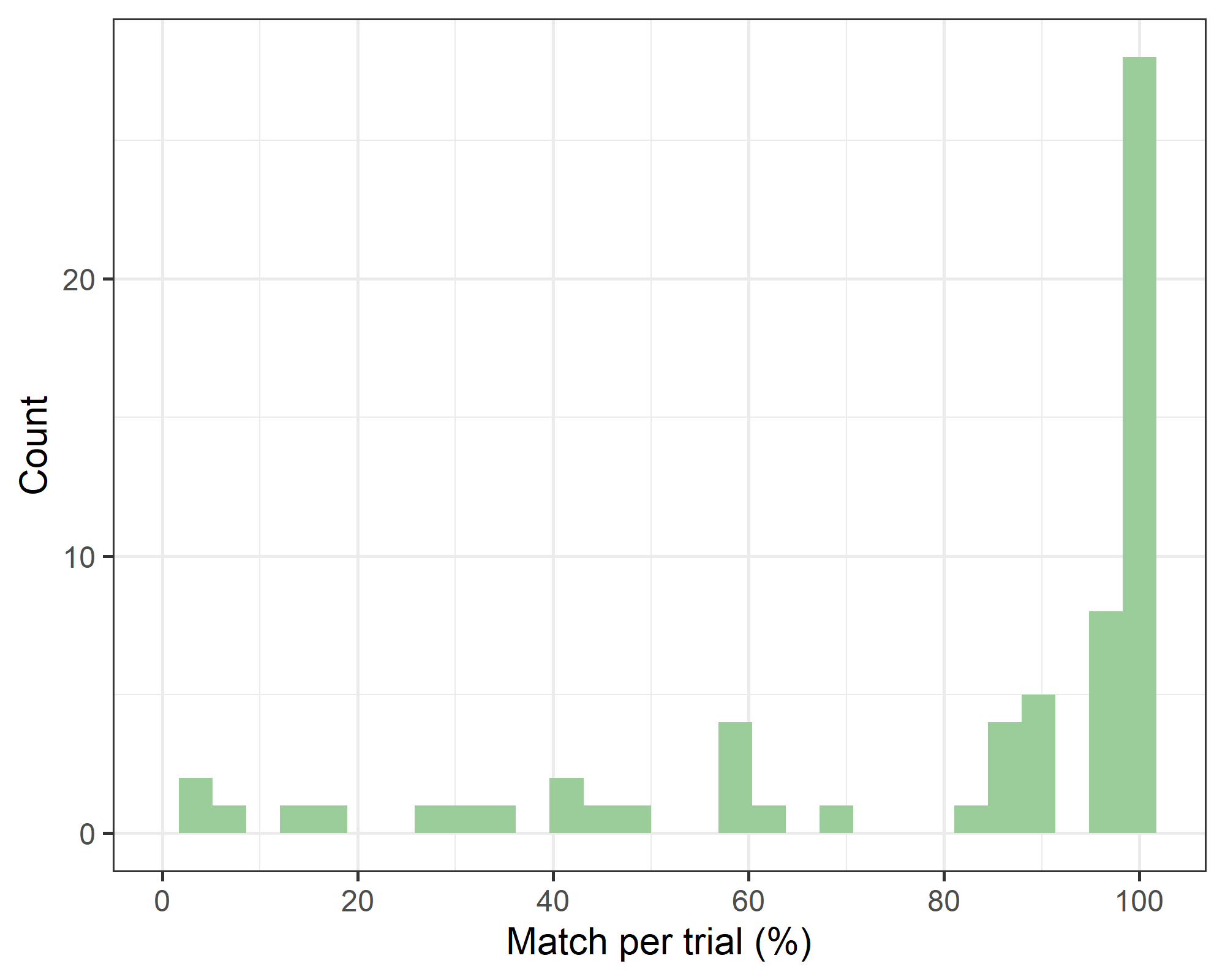
### c) Overall statistics for statistics type

| **Match** | **n** | **percent** | **Valid percent** |
| --- | --- | --- | --- |
| FALSE | 26 | 3 | 3 |
| TRUE | 743 | 89 | 97 |
| Missing | 63 | 8 |  |
| Total | 832 | 100 | 100 |

The table shows the overall percent of matches for every table row. An exact 95% confidence interval for the match percentage is 87% to 91%.

## Differences by numerical statistic

Here we look at the difference in the statistics as numbers. We assume a match where the two numbers differ by less than 0.01.



Some mismatches were caused by table columns that were not aligned (PMC7424488). This could be improved by a more complex matching that used the statistics in the rows to better align the manual and algorithm data.

One large error was for a table which only gave percentages for some rows, but the algorithm recalculated the numerator which caused multiple mismatches in the statistics as the calculated numerators were compared with the percentages (PMC6966838). These mismatch errors are not a concern for the model which will use the numerators.

#### Summary statistics for close matches in statistics

| **Statistic type** | **Statistic** | **n / N (%)** |
| --- | --- | --- |
| confidence interval | First | 4 / 4 (100.0) |
| Second | 0 / 4 (0.0) |
| continuous | First | 570 / 621 (91.8) |
| Second | 568 / 613 (92.7) |
| median | First | 129 / 151 (85.4) |
| Second | 124 / 142 (87.3) |
| min\_max | First | 2 / 2 (100.0) |
| Second | 2 / 2 (100.0) |
| numbers | First | 61 / 67 (91.0) |
| Second | 54 / 55 (98.2) |
| percent | First | 1110 / 1321 (84.0) |
| Second | 1059 / 1140 (92.9) |

The table shows the number of close matches of an under 1% difference. The statistics are given by the statistic type and whether it was the first or second statistic in the table cell (e.g., for continuous data it would be the mean first and standard deviation second). These results ignore missing data and just look at where there were two statistics to compare.

The results were somewhat worse for the median and percentages.

#### Table as per above but only where row and column numbers match

| **Statistic type** | **Statistic** | **n / N (%)** |
| --- | --- | --- |
| confidence interval | First | 4 / 4 (100.0) |
| Second | 0 / 4 (0.0) |
| continuous | First | 497 / 499 (99.6) |
| Second | 491 / 499 (98.4) |
| median | First | 115 / 115 (100.0) |
| Second | 110 / 110 (100.0) |
| numbers | First | 56 / 56 (100.0) |
| Second | 48 / 48 (100.0) |
| percent | First | 872 / 928 (94.0) |
| Second | 830 / 835 (99.4) |

This table only compares baseline tables where the number of rows and columns for the algorithm and manually entered data were the same. The aim is to reduce the number of misaligned results. The table again shows the number and percent of matches within 1%.

The percentage of matches is clearly better than above, confirming that the mismatching of table rows is the major issue.

Some differences were because the algorithm changed the two statistics of a percentage alone into a number, as that (combined with the sample size) is required for the t-statistic (e.g., PMC7568358).

# Sample size

This section examines differences in sample sizes.

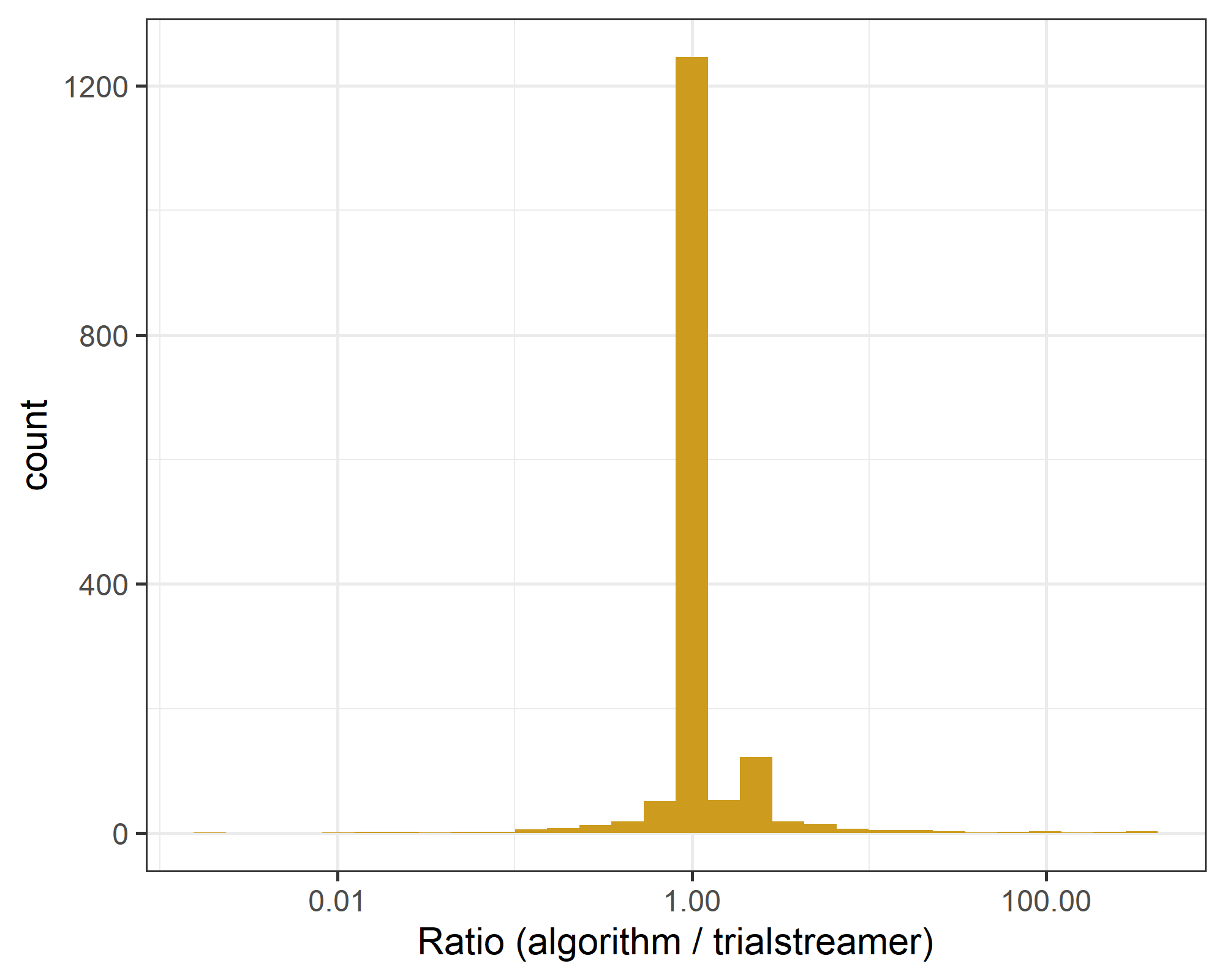
| **n** | **median** | **q5** | **q95** | **min** | **max** |
| --- | --- | --- | --- | --- | --- |
| 181 | 0 | 0 | 0 | -18 | 234 |

The table shows the median difference in sample size, the 5% to 95% percentile of the difference, and the range. Sample size was almost perfectly estimated by the algorithm.

The biggest difference was due to a table with a relatively complex format that reported stratified results that were badly aligned (PMC7028836).

# Trialstreamer sample size

I compared the total sample size in my algorithm with that estimated by the *trialstreamer* program [1] for the 1,684 trials which had estimated sample sizes from both approaches.



This is the difference in the total sample size that the algorithm extracted from the baseline table and the number randomised estimated by *trialstreamer*. The number of comparisons was 1596.

The x-axis is on a log-scale (base 10). There were 73% of studies where the difference was within plus/minus 10%.

Large differences were explained by:

* where *trialstreamer* gave the number of clusters in the randomised trial, whereas the algorithm gave the much larger number of participants (PMC7690135).
* because *trialstreamer* used the year from the abstract rather than the number randomised (PMC4852093).
* because the baseline table used a subset of the full sample due to missing data or drop-out and the authors wrote the full sample in the abstract (PMC6746080).

# Reference

1 Marshall IJ, Nye B, Kuiper J, *et al.* Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association* 2020;**27**:1903–12. doi:[10.1093/jamia/ocaa163](https://doi.org/10.1093/jamia/ocaa163)