**Integrating multiple rolling recruiting lists**

William Beasley, David Bard, Thomas Wilson

**Background:**

Technical challenges arose while continually recruiting participants for three years. List of individuals were provided by collaborating state organizations that feel the study’s results will inform their policy and advance their mission statement. There were three primary atypical challenges.

1. Families from multiple lists from *a single organization* need to be matched every few months, without a common family numeric identifier.
2. Families from *multiple organizations* need to be matched every few months without a common numeric identifier.
3. Each new list needs to be merged with the live recruiting database in a way that stale information is overwritten, but manually-entered information is not.

This is in addition to the challenges typically encountered during clinical recruitment:

1. Individuals who previously declined should not be re-contacted when a new source contains different contact information.
2. Decision processes must be automated, in order to feasibly process 10,000 potential recruits in each list.
3. De-duplication thresholds should not be too liberal. Too many false-positives produce an imbalanced sample that underrepresents and underserves subgroups.

**Methods:**

Recruiting lists were received as CSV (comma separated value) text files, and were transferred to a SQL Server database to improve security and robustness during subsequent data manipulation. Record matching and deduplication were conducted with R, an open-source software system designed for manipulation and data analysis. Once complete, records were imported into REDCap, a system for clinical databases that is free to OUHSC Pediatrics researchers. Recruiters used REDCap to select individuals and record the outcomes.

After exact matching strategies were attempted to link records from different sources, approximate matching (i.e., ‘fuzzy matching’) were attempted. This allowed the linking process to be tolerant of small, but less meaningful, differences such as transposed letters. Relative frequencies were considered so that it was less likely to match records with a common name like ‘Wilson’, than it was for less common names like ‘Beasley’ or ‘Bard’.

**Results and Conclusions:**

Subsets of the data were manually inspected and used to train the algorithm (e.g., tune parameters of the approximate matching routine). After several months of new lists and algorithm adjustments, managing the process required less time, and recruiting participants became more efficient.

Removed to save space:

1. Ineligible recruits should not be contacted, so that the study is less intrusive and more efficient.