Development of PEDSNet Data Extraction Code Using GitHub

Aaron N Browne, Jeffrey Pennington, Charles Bailey, PhD
Children's Hospital of Philadelphia, Philadelphia, PA

Abstract

The Pediatric Learning Health System (PEDSNet) is in the process of standardizing and aggregating electronic medical records from its eight member hospitals. The members of PEDSNet are using GitHub to collaboratively work towards this goal. A structure and workflow have been developed and the system is in active, effective use.

Introduction and Background

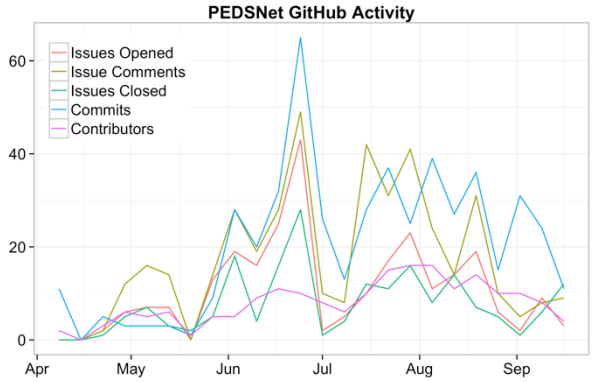
Research studies that aggregate and standardize electronic medical records across multiple institutions have great potential. However, such efforts routinely suffer from data quality issues. We at the Pediatric Learning Health System (PEDSNet) chose to develop our methods out in the open, from the beginning. The GitHub platform has allowed us to collaborate rapidly and effectively, addressing many potential data quality problems before they became embedded in practice. We present the structures and collaboration workflow we developed as well as metadata showing the extent of our collaboration thus far.

Methods

We developed a structure and workflow for the conversations, documentation, and code that would be stored on GitHub, including granular access permissions, based on best practices from the open source software development community. The data we present were collected and analyzed using the GitHub API and small scripts written in Python and R, which are available on GitHub[1].

Results

A PEDSNet organization was created on GitHub, and teams within that on a per-site basis to organize permissions. Repositories have been created to meet each of the following needs: eight for site-specific extraction code, two for documentation (data models and Data Coordinating Center (DCC) planning), three for DCC operational code, and one for query code distribution. The workflow we developed starts with any organization member opening an issue, progresses through discussion via comments, and finishes with code or documentation changes being committed. At the discretion of the implementer, the original issue may be closed and one or more implementation issues opened to track the changes.



The figure shows the number of PEDSNet organization contributors, issues opened, comments on issues, issues closed, and changes committed (commits) as measured on a weekly basis (average repository age is 18 weeks). Respectively, the averages and standard deviations of the displayed variables are 8.0±4.6, 11±10, 17±14, 7.8±6.9, and 21±16. In total, the organization has 56 contributors, 259 issues, 415 issue comments, 186 closed issues, 645 commits, and 2,846,432 lines of code (5,439,682 added and 2,593,250 removed). All correlations between the displayed variables were high ($r > 0.68$) except for the correlations from contributors to issues opened and to issues closed, where the correlations where moderate ($r > 0.44$).

Discussion

GitHub has been highly effective at facilitating collaboration on data extraction code and documentation across the PEDSNet sites. The figure shows generally increasing use, with periodic local maxima and minima aligning with major network deliverables at the ends of May and June and throughout July and August. The high correlations between the variables shows that our workflow has been functioning as intended, with issues driving conversations and code changes, although observation over a longer time period is needed. Generally, the amount and ease of collaboration through GitHub has been impressive and we look forward to its continued use.

---

[1] https://github.com/aaron0browne/pedsnet-github-amia