Promoting Data Quality in a Clinical Data Research Network Using GitHub

Aaron N. Browne[1], Jeffrey W. Pennington[1], L. Charles Bailey, MD, PhD[2]

[1]Center for Biomedical Informatics, Children's Hospital of Philadelphia, PA; [2]Department of Pediatrics, Children's Hospital of Philadelphia and Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

Abstract

The Pediatric Learning Health System (PEDSnet), a multi-institutional pediatric clinical data research network, is standardizing and aggregating electronic health records from its eight member hospitals. The members of PEDSnet are using GitHub to collaboratively and transparently develop methods that produce high-quality data.

Introduction

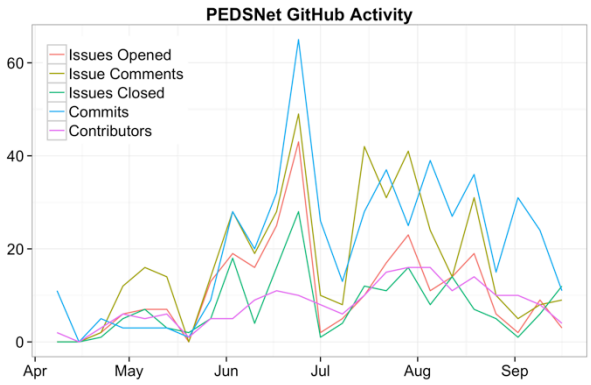
Research studies that aggregate and standardize electronic health records have great potential. However, such efforts routinely suffer from data quality issues. The Pediatric Learning Health System (PEDSnet) informatics investigators chose to develop our methods transparently and in ways that facilitate sharing. The GitHub platform has allowed us to collaborate rapidly and effectively, addressing many potential data quality problems before they became embedded in practice. It also increased efficiency by facilitating sharing of code and documentation across teams. We present the structures and workflow we developed as well as metadata measuring the extent of our collaboration.

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 were 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. Custom issue labels, separate implementation issues, and implementation branches were all used as needed.



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, as of September 2014, the organization has 56 contributors, 259 issues, 415 issue comments, 186 closed issues and 645 commits. 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 ($r > 0.44$). Lines of code added and removed are not shown because these values were skewed by the inclusion of proprietary XML-based-format files (such as docx, dtsx, etc), which have many lines of formatting "code".

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. 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