**Development of PEDSNet Data Extraction Code Using GitHub**

**Aaron N Browne, BA1**

**1Center for Biomedical Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA**

**Abstract**

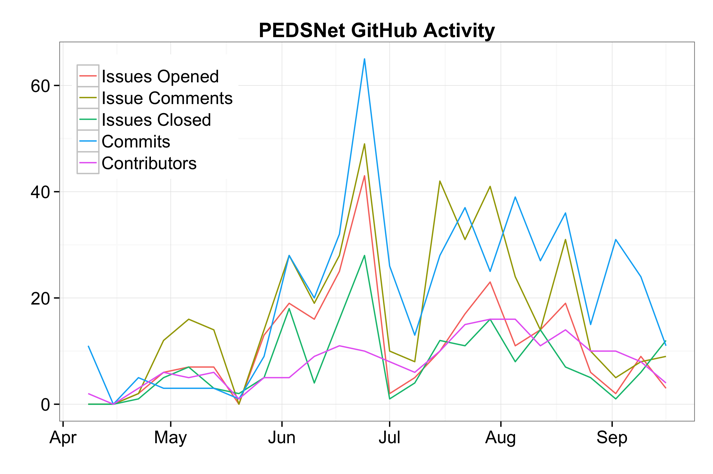
*The Pediatric Learning Health System (PEDSNet) is in the process of standardizing and aggregating electronic medical records from its eight member hospitals. The members of PEDSnet are collaborating*

**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]](#footnote-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. Each of these five values show correlation with the others. All correlations were high at r > 0.68 except for the correlations between contributors and issues opened and closed, where the correlations where moderate at r > 0.44

**Discussion**

Similar efforts have routinely suffered from data quality issues; between site differences in source data models and data extraction methods are repeatedly cited as a main cause of poor data quality. We offer a model for establishing highly collaborative and transparent methods development practices early in the data collaboration effort using GitHub as a communication, project management, and code-sharing tool. By increasing shared decision-making and distributing effort asynchronously, our model can increase method consistency across sites and therefore data quality even before it would traditionally be assessed. Such improvements early in the collaboration have the potential to save significant effort in its later stages. We present a case study of the implementation of our model in the Pediatric Health Learning System Clinical Data Research Network.

**Introduction**

Retrospective Electronic Health Record (EHR) research is a rapidly growing study methodology across a wide range of health science disciplines. By extracting and analyzing the millions of clinical care data points already available in established Electronic Medical Record (EMR) systems, these studies have the potential to rapidly increase the clinical care evidence base. Fields as diverse as comparative effectiveness, pharmaceutical and procedure risk, health care cost and geographic variation, care of special populations, and personalized medicine are taking advantage of this cost-effective methodology. In addition to providing insight on specific research questions, review of clinical care records can inform system-wide changes in payment models, value-based care methods, and research recruitment strategies.

In their efforts to harness this potential, many studies are moving beyond the data available to them locally and attempting to combine datasets across disparate care sites. Although this increase of scale does offer

1. [↑](#footnote-ref-1)