Distributed Regression Analysis in a Distributed Health Data Network

Jessica M. Malenfant, MPH; Qoua L. Her, PharmD, MSPharm, MSc; Sarah Malek, MPA; Yury Vilk, PhD; Sengwee Toh, ScD

Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA

33rd INTERNATIONAL CONFERENCE ON PHARMACOEPIDEMIOLOGY & THERAPEUTIC RISK MANAGEMENT, Palais de Congres de Montreal, Montreal, Canada, August 26-30, 2017

ABSTRACT

Background: Distributed health data networks use distributed databases for efficient, privacy-protecting, and effective public health research and surveillance activities. Distributed regression analysis (DRA) is a novel analytic method that does not require transferring of patient-level data in multi-database studies but produces results statistically equivalent to those from pooled patient-level data analysis. The execution of DRA has been largely manual and labor-intensive. We describe a new approach to conduct automated DRA in the FDA's Sentinel system, a distributed network using multiple electronic health data sources for medical product safety monitoring.

Objective: Implement a method within the existing PopMedNet™ (PMN) open-source platform used in Sentinel to allow automated, iterative, privacy-protecting, and scientifically accurate DRA in a real-world setting.

Methods: The project had 2 work streams: (1) develop DRA analytic code in SAS for multivariable-adjusted regression models and (2) enhance PMN to process DRA automated communication cycles within the distributed data network. We developed a new capability in PMN to enable the analysis center to (1) automatically aggregate site-specific intermediate statistics to compute or update the parameter estimates, which are returned to the data partners for subsequent iterations, and (2) to allow this iterative process to continuously refine the statistics until the model converges. The main outcome of interest was confirmation of analytic code accuracy and execution of DRA in a real-world setting. The DRA analytic code was validated against test data using results from pooled patient-level data analysis as a benchmark. PMN automation was tested internally and with external data partners.

Results: PMN software development was an iterative process where the implementation ensured that the functionality developed within the PMN code base would not impact existing Sentinel workflows or system functions. We developed and validated PMN's ability to perform regression analysis using only summary-level intermediate statistics and produce statistically equivalent regression parameters as pooled individual-level data analysis.

Conclusion: This work can be leveraged in the future for DRA in Sentinel and other networks. The functionality is agnostic to statistical software and can be extended to R and other software.

Funding: Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

OBJECTIVE

- Implement a system to conduct distributed regression analysis (DRA) within the existing PopMedNet™ (PMN) open-source platform
- DRA key development features:
  - Automated
  - Iterative
  - Privacy-protecting
  - Scientifically accurate
  - Demonstrated use in a real-world setting

BACKGROUND

- Distributed databases enable efficient, privacy-protecting, and effective public health research and surveillance activities
- DRA is a novel analytic method that **does not require transferring of patient-level data**
- DRA produces results statistically equivalent to pooled patient-level data analysis
- To date, DRA has been largely manual and labor-intensive

METHODS

The project had two work streams:

- **Work stream 1:** Develop DRA analytic code in SAS for multivariable-adjusted regression analysis
- **Work stream 2:** Enhance PMN to process DRA automated communication cycles within the distributed network

Extend PMN to enable the Analysis Center to:

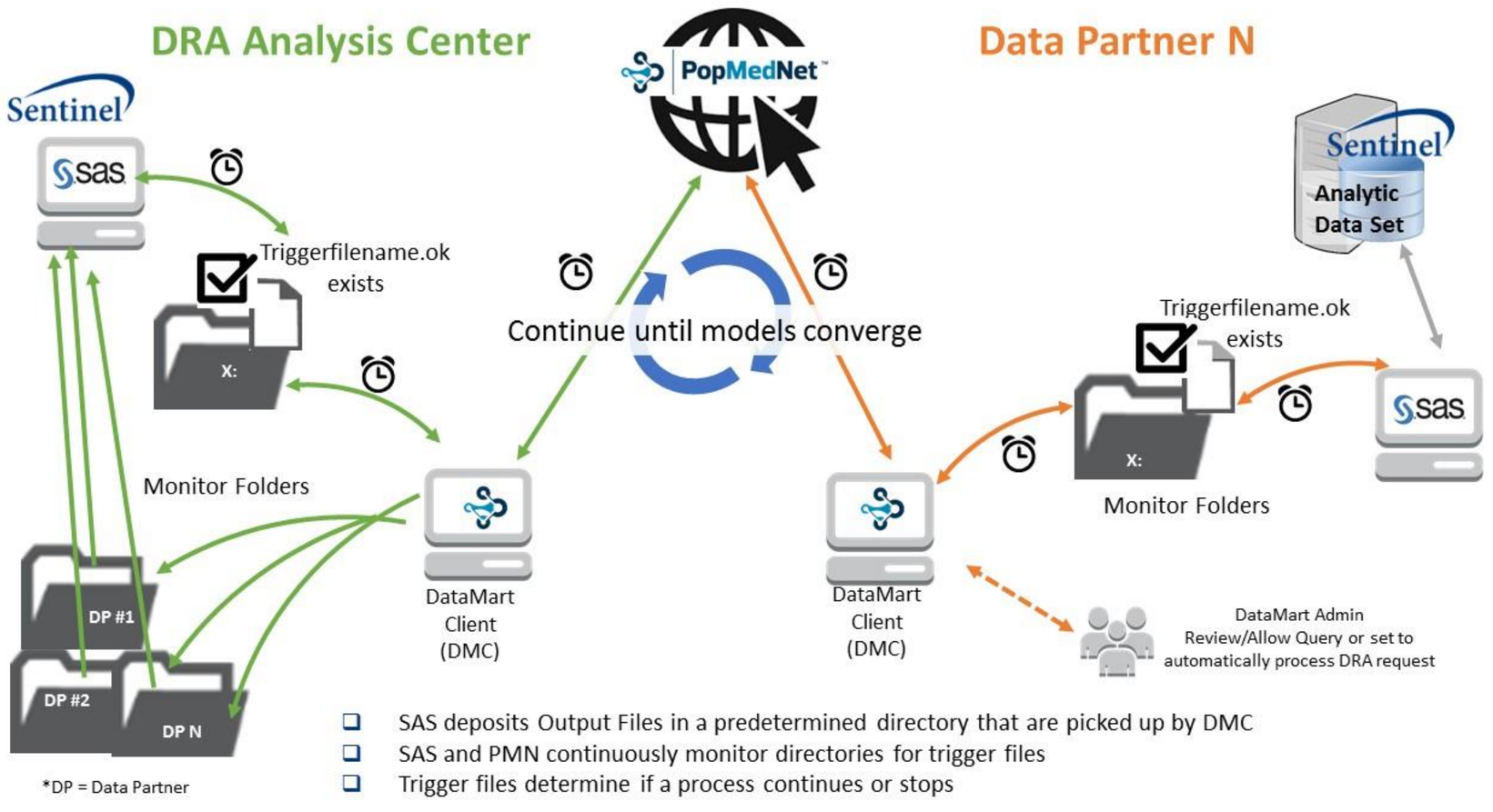
- Automatically aggregate site-specific intermediate statistics to compute regression parameter estimates, which are returned to the data partners for subsequent iterations
- Allow this iterative process to continuously refine the regression statistics until the model converges

Implement and validate DRA in real-world setting:

- DRA analytic code was validated against test data
- PMN automation was tested internally and with external data partners

IMPLEMENTATION & RESULTS

- PMN software development was an iterative process where the implementation ensured that the functionality developed would not impact existing workflows or system functions
- We developed and validated PMN's ability to perform DRA using only intermediate statistics and produce statistically equivalent regression parameters as pooled individual-level data analysis
- Iterations driven by trigger files that indicate the start or end of a process, used by SAS and PMN
- Implementation details illustrated in the following diagrams describe the analytic and file transfer processes



CONCLUSION

- The DRA analytic code demonstrated reliable, accurate results when applied to test data sets; next steps include testing with Sentinel production data
- Trigger files created and processed by the DRA analytic code and PopMedNet drive the automation and integration for successful DRA
- This work can be leveraged in the future for DRA in Sentinel and other distributed health data networks
- The functionality is agnostic to statistical software and can be extended to R and other software

DISCLOSURE STATEMENT

- Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.
- The authors have no relationships to disclose.