Integrating R efficiently to allow secure, interactive analysis within a clinical data warehouse

Daniel W. Connolly[1,*], Bhargav Adagarla[1] , John Keighley[1] , Lemuel R. Waitman[1]

1. Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas

*Contact author: dconnolly@kumc.edu

Keywords: Data Warehouse, Security, Biomedical Informatics

HERON[1] is a clinical data repository with about a half a billion facts, linking hospital medical records and clinic billing systems with our tumor registry and biospecimen repository and national databases such as the Social Security death index. It is built on i2b2, an NIH-funded scalable informatics framework[2]. Users can interactively answer questions such as "how many patients meet my study criteria?" and use analysis plug-ins to visualize the potential study cohort's demographic characteristics. Bulk export of data for off-line analysis is allowed but only after explicit approval by a governance structure that safeguards patient privacy. We aim to package a range of statistical analysis methods for secure, on-line analysis within i2b2.

On top of a star-schema database and a middle tier web services *hive of cells*, i2b2 provides a web based user interface. Previous work developed *RECell* [3] to integrate *R* into the i2b2 architecture and a Kaplan Meier analysis plug-in that relays patient data as XML to the *RECell.* The *RECell* then transforms data as required by the R **survival** package, and invokes *R* to produce a plot which is fetched and displayed by the plug-in. Rather than sending all the data via the web client and serializing/parsing it several times, our approach, *rgate*, connects R directly to the database, using the **DBI** and **ROracle** packages, performing the data transformation in *R*, and proceeds with the survival package as above.

Authority flows from an oversight committee (hospital, clinics, and university), which grants users authority to view data and grant operators authority to run the service; the operator grants *rgate* authority to query the database via a configuration file; users present login credentials as proof of their authority to view data; the plug-in relays this authority to *rgate* in the form of a session identifier.

The *R* code invoked by *rgate* is split between a trusted deid.R broker and an untrusted analysis.R module. The trusted broker module has a function that takes (1) database login credentials and (2) a patient set identifier and returns an facet object[4] that will only query attributes of those patients. Only this facet object is given to the untrusted analysis module. This isolates the statistician from security and governance issues to ensure their statistical analysis modules do not cause users to exceed the authority given to them by the oversight committee.

References

[1] Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing Observations from Electronic Medical Record Flowsheets in an i2b2 based Clinical Data Repository to Support Research and Quality Improvement. AMIA Annu Symp Proc. 2011;2011:1454-63. Epub 2011 Oct 22.

[2] Murphy SN et al. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc.* 17, 124–130.

[3] Segagni D, e. a. (2011, May). R engine cell: integrating R into the i2b2 software infrastructure. *J Am Med Inform Assoc.* 18(3), 314–7.

[4] Miller, M. S. (2006). Robust Composition: Towards a Unified Approach to Access Control and Concurrency Control. http://www.erights.org/talks/thesis/.