

# Extending an I2B2-based Clinical Data Repository with the R Statistical Platform

Daniel W. Connolly, Lemuel R. Waitman

University of Kansas Medical Center, Kansas City, Kansas

HERON<sup>1</sup> includes the usual i2b2 cohort query facilities on a largely self-service basis as well as bulk export of data for off-line analysis with approval by an oversight committee. We aim to support richer analysis without the need for bulk export.

We have adapted the survival analysis plug-ins from the **R Engine Cell** work<sup>2</sup> to address scalability limitations and provide secure extensibility. Our approach, **rgate**<sup>3</sup>, has one privileged database access module, which uses R and the ROracle package to run SQL queries directly against our data warehouse, along with any number of unprivileged analysis scripts. The analysis scripts for our survival plug-ins use the R survival package for plotting.

**RStudio Server**<sup>4</sup> provides an integrated development environment (IDE) for R over the web. HERON integrates with RStudio via an **R Data Builder** plug-in (Figure 1). Like the i2b2 timeline, this plug-in takes a patient set and a set of concepts as input (Figure 2). The patient set identifier and concept paths, along with a filename of the user's choosing, are sent via rgate to the R Data Builder analysis script, which queries the data warehouse and saves the results on the server. Then the user logs in to RStudio Server and uses the R readRDS() function to load the data for further analysis (Figure 3). Like i2b2, and unlike bulk export, this approach is designed to keep the data on the server in our data center.

## Applications

The survival analysis plug-ins are available to the entire HERON user-base. In addition to typical applications such as 5-year cancer survival plots, students have used it to visualize 28-day sepsis outcomes.

The HERON study team has piloted the R Data Builder and RStudio Server in work-in-progress on a quality-improvement investigation into medication timing.

We have not established governance policies for use of RStudio Server by the general HERON user-base. While HERON users execute a system access agreement that prohibits screenshots, printing, and other data export, the risks associated with RStudio Server are significantly higher than with I2B2 plug-ins. For example, R includes facilities to email a data set, and we have not yet put in place any technical limitation on such facilities.

---

<sup>1</sup> Waitman, Lemuel R et al. "Expressing Observations from Electronic Medical Record Flowsheets in an i2b2 based Clinical Data Repository to Support Research and Quality Improvement." *AMIA Annual Symposium Proceedings* 2011: 1454.

<sup>2</sup> Segagni, Daniele et al. "R Engine Cell: integrating R into the i2b2 software infrastructure." *Journal of the American Medical Informatics Association* 18.3 (2011): 314-317.

<sup>3</sup> Connolly, Daniel W et al. "Integrating R efficiently to allow secure, interactive analysis within a clinical data warehouse." R User Conference, Nashville TN, June 2012

<sup>4</sup> Racine, Jeffrey S. "RStudio: A Platform-Independent IDE for R and Sweave." *Journal of Applied Econometrics* 27.1 (2012): 167-172.

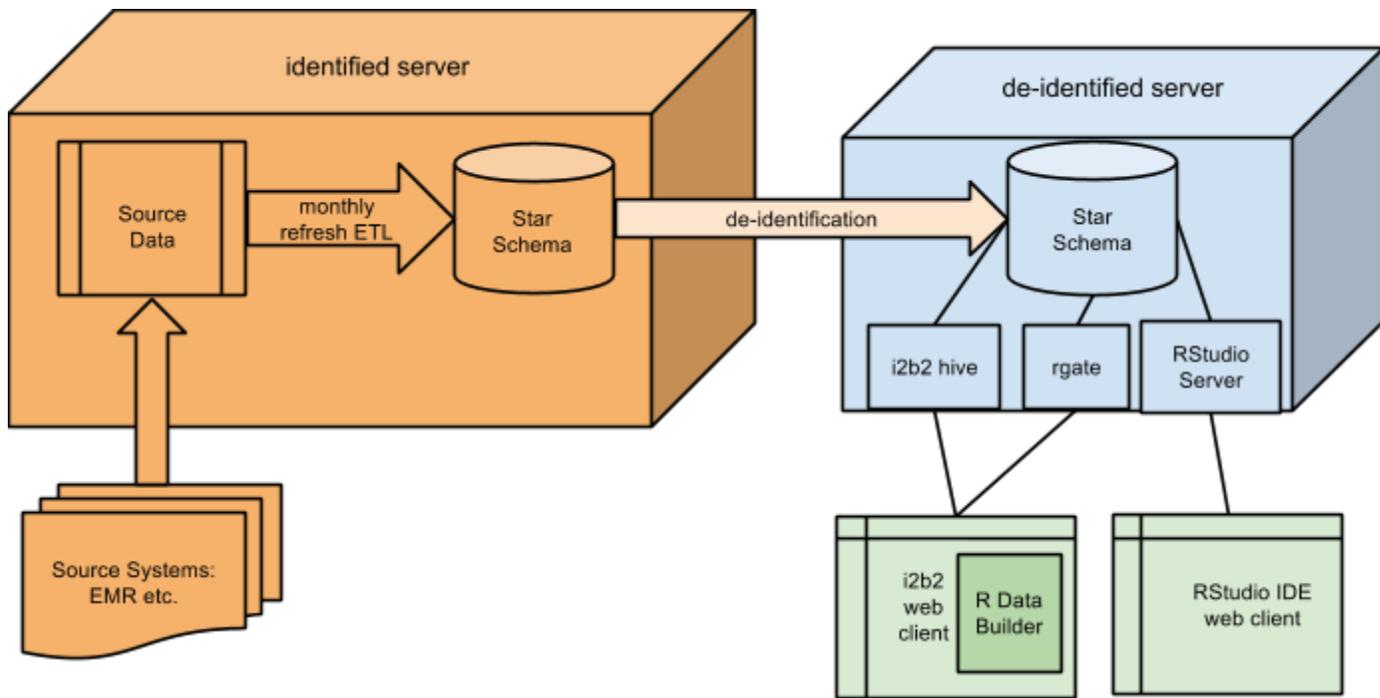


Figure 1: HERON Architecture.

User: Dan Connolly Find Patients | Analysis Tools | Message Log | Help |

**DataFrameBuilder**

## R Data Builder

**Note:** Access to [RStudio Server](#) is currently limited to members of the HERON study team.

Patient Set: Lung an-Never [@10:16:52 [2-14-2013] [dconnolly] [PATIENTSET\_11]

The [Running a Query](#) section shows how to create one.

Patient Data: *Birthdate, sex, vital status, race, etc. are automatically included.*

Other Observations: Lung and Bronchus [6,911 facts; 6,842 patients]  
Never [571,137 facts; 141,196 patients]

Observations from Query: Lung an-Never [@10:16:52 [2-14-2013] [dconnolly]

File: /home/dconnolly /heron/  .Rda

**Build R Data** **Test Build R Data**

```
items <- readRDS('/home/dconnolly/heron/lung-never-smoked.Rda')
str(items)
```

	name	n.patient	n.obs
1	Lung and Bronchus	167	171
2	Never	167	1182

---

**Plugins**

Detailed List View Category: ALL

- Analysis of KUH Tumor Registry data using the R survival library
- Data Builder for RStudio**  
Build R data frame from patient set and variable concepts
- Multi-Cohort Survival Analysis

Figure 2: R Data Builder Plug-in User Interface

File Edit View History Bookmarks Tools Help

RStudio

https://.../rstudio-server/

DWC MyLibrary KU KUMC BMI KU DanStatus KU trac+= KUMC Fading i2b2 Web Client KU HERON Research Dat... Read Now

Diigo Bookmark Highlight Read Later Options Go premium!

File Edit Code View Project Workspace Plots Tools Help dconnolly | Sign Out Project: (None)

cancer x 171 observations of 11 variables

	start.date	end.date	patient.num	encounter.num	code	modifier	instan
1	2000-10-04	2000-10-04	15975	22984831	SEER_SITE:22030	@	1
2	2011-11-17	2011-11-17	17864	22999655	SEER_SITE:22030	@	1
3	2002-04-10	2002-04-10	22168	22977762	SEER_SITE:22030	@	1
4	2008-09-09	2008-09-09	23095	22990931	SEER_SITE:22030	@	1
5	2009-04-19	2009-04-19	62857	22991689	SEER_SITE:22030	@	1
6	2010-07-09	2010-07-09	63172	22998747	SEER_SITE:22030	@	1
7	2004-12-10	2004-12-10	87124	22982206	SEER_SITE:22030	@	1

Workspace History

Load Save Import Dataset Clear All

Data

cancer 171 obs. of 11 variables

Values

items list[5]

Functions

cohort.summary(q)  
demographics(x, ...)  
demographics.clean(raw)

Console ~

```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

Loading required package: ROracle
Loading required package: DBI
> items <- readRDS('/home/dconnolly/heron/lung-never-smoked.Rda')
> source('projects/rgate/rgate/deid_test.R')
> cohort.summary(items)
Error: could not find function "cohort.summary"
> source('projects/rgate/rgate/deid.R')
> cohort.summary(items)
      name n.patient n.obs
1 Lung and Bronchus    167    171
2      Never          167   1182
> summary(items$patient$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00  57.00   66.00  65.39  76.00   89.00
> cancer <- q.d(items, 'Lung')
> View(cancer)
> plot(hist(items$patient$age))
>
```

Files Plots Packages Help

Zoom Export Clear All

**Histogram of items\$patient\$age**

Frequency

items\$patient\$age

Figure 3: Analyzing data from I2B2 in RStudio.