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

Daniel W. Connolly, Bhargav Adagarla, John Keighley, Lemuel R. Waitman

University of Kansas Medical Center
Kansas City, Kansas

This project is supported in part by NIH grant UL1TR000001 and NSF Award CNS-1258315
HERON Research Support Goals

Clinical Data Repository supports:

- Cohort Discovery
  - prospective trials: feasibility
  - retrospective studies: data use
- Hypothesis Generation
  - explore data
  - summarize
  - visualize


informatics.kumc.edu
HERON helps Investigators Identify Cohorts using I2B2

HERON Architecture: Where Informatics and Governance Roles Meet

- Data from Epic Clarity database (> 7,000 tables & 60,000 columns)
- Transformed into an I2B2-compatible schema. Then, de-identified, and loaded on a separate database server to be accessed by I2B2.
- De-identified data used by I2B2 is deemed non-human subjects research by our institutional review board
Richer Analysis Without Bulk Export

• HERON includes i2b2 query on a largely *self-service* basis.

• Bulk export of data for off-line analysis involves approval by an oversight committee.

• Aim: support *richer analysis without the need for bulk export.*
Segagni D, Ferrazzi F, Larizza C, Tibollo V, Napolitano C, Priori SG, Bellazzi R.

R Engine Cell adds Survival Plotting
Programs and Products

• Program: works for the programmer
• Product: works for the customer
• Sharing programs is useful even though they’re not products yet, or ever!

“Programs must be written for people to read, and only incidentally for machines to execute.”

Hal Abelson, *Structure and Interpretation of Computer Programs*
Integrating the R Engine Cell with HERON for Cancer Research

Issues:

- **Clinical Domain**
  - cardio vs. cancer
  - start at birth vs start at diagnosis
  - stratification: gender vs. stage

- **Version Skew**
  - RE Cell: I2B2 version 1.4
  - HERON: I2B2 version 1.6

- **Architecture...**

Photo credit: Christopher Harshaw
Efficiency, Scalability: R Engine Cell Data Path

CRC Cell sends back to the **plug-in** an **XML response containing the requested data** (extracted from the i2b2 datawarehouse).
Efficiency, Scalability:
rgate connects R to Oracle directly

Like the CRC cell, rgate calls the PM cell to validate authorization.
R Engine Cell approach to R Integration: Kaplan Meier jar application

R Code Generation in KMAnalysis.java:

```java
Integer[] statusInteger = (Integer[]) status.toArray(new Integer[status.size()]);
StringBuffer statusStr = new StringBuffer();
statusStr.append("status<-c(";
for(int i=0;i<statusInteger.length;i++){
    statusStr.append(statusInteger[i].intValue());
    if(i!=(statusInteger.length-1))
        statusStr.append",");
}
statusStr.append(")");

re.eval("data=data.frame(time,status,gender)");
re.eval("names(data)=c('time','status','gender')");
re.eval("setwd("+resultFolder+")");
re.eval("library(survival)");
re.eval("fit <- survfit(Surv(data$time, data$status) ~ gender, data)");
```
R Engine Cell approach to R Integration: Kaplan Meier jar application

R Code Generation in KMAnalysis.java:

```java
...  
    Integer[] statusInteger = (Integer[]) status.toArray(new Integer[status.size()]);
    StringBuffer statusStr = new StringBuffer();
    statusStr.append("status<-c(");
    for(int i=0;i<statusInteger.length;i++){
        statusStr.append(statusInteger[i].intValue());
        if(i!=(statusInteger.length-1))
            statusStr.append(", ");
    }
    statusStr.append(");
    ...

    re.eval("data=data.frame(time,status,gender) ");
    re.eval("names(data)=c('time','status','gender') ");
    re.eval("setwd("\""+resultFolder+"\"") ");
    re.eval("library(survival) ");
    re.eval("fit <- survfit(Surv(data$time, data$status) ~ gender) ");
...```

biostatistics, R
Separation of Concerns in rgate: R code goes in .R files

Analysis is written in the language of statisticians:

```r
##' km_analysis -- Kaplan Meyer analysis from i2b2 observations
library(ROracle)
acct = db_config()
patient.set.survival <- function(concept.paths, patient.set.id, 
    web.folder, filename) {
    conn <- dbConnect(Oracle(), acct$username, acct$password, access)
    sql <- paste("select ", concept.paths$event, ", " panel

    , to_char(f.start_date, 'YYYY-MM-DD HH24:MI:SS') start_date
    , pset.patient_num
    , cd.name_char
    , cd.concept_cd
    from blueherondata.observation_fact f, ...")
    data = transform.observations(dbGetQuery(conn, sql))
    fit <- survfit(Surv(data$time, data$status) ~ concept.paths$stratum, data)
    png(paste(web.folder, filename, sep='/'))
    plot(fit, xlab="Time (Years)", ylab="Survival probability")
    dev.off()
}
```
Separation of Concerns in rgate: R code goes in .R files, but...

How well does the R code behave when the author is not there?:

```r
##' km_analysis -- Kaplan Meyer analysis from i2b2 observations
library(ROracle)
acct = db_config()
patient.set.survival <- function(concept.paths, patient.set.id, 
  web.folder, filename) {
  conn <- dbConnect(Oracle(), acct$username, acct$password, access)
  sql <- paste("select ", concept.paths$event, ", " panel
                , to_char(f.start_date, 'YYYY-MM-DD HH24:MI:SS') start_date
                , pset.patient_num
                , cd.name_char
                , cd.concept_cd
  from blueherondata.observation_fact f, ...")
data = transform.observations(dbGetQuery(conn, sql))
fit <- survfit(Surv(data$time, data$status) ~ concept.paths$stratum, data)
png(paste(web.folder, filename, sep='/'))
plot(fit, xlab="Time (Years)", ylab="Survival probability")
dev.off()
```
Object Capability Discipline supports the Principle of Least Authority

Memory safety and encapsulation\(^1\)
+ Effects only by using held references\(^2\)
+ No powerful references by default\(^3\)

Reference graph \(\equiv\) Access graph
Only connectivity begets connectivity

Natural Least Authority

OO expressiveness for security patterns

M. Miller, C. Morningstar, B. Frantz; "Capability-based Financial Instruments"; Proceedings of Financial Cryptography (Springer-Verlag); 2000

\(^1\) closure inspection is not safe: \texttt{environment}(function), \texttt{as.list} (function)
\(^2\) \texttt{plot}(fit) implicitly uses results of \texttt{png}(paste(web.folder, filename))
\(^3\) R global environment most likely includes lots of powerful references
rgate Security Architecture:
km_analysis.R can only read patient set

Attenuated patient data access:
```r
##' km_analysis -- Kaplan Meyer analysis from i2b2 observations
library(survival)
run_analysis <- function(patient.set, folder, filename, progress,
                         paths, title, xmax) {
  obs.db = observations(patient.set, unlist(paths))
  progress(paste("query returned", nrow(obs.db), " observations."))

  data <- db2km(obs.db, paths)
  progress(paste("db2km resulted in ", nrow(data), " data points for plotting."))

  survplot(data, title, folder, xmax, filename)
  progress(paste("KM plot stored in", filename, " in", folder))
}
```
Interactive R statistical visualization in HERON Clinical Data Repository
Exploring Breast Cancer comorbidities: Obesity, Diabetes

HERON brings together diabetes diagnosis and BMI from hospital EMR with cancer staging from tumor registry and vital status from the U.S. SSA death index.
R Survival Plug-in in Regular use by HERON Users

- Patient privacy, institutional liability
- Python, SQL, HTML, JavaScript
- Biostatistics, R
- Cancer prevention, treatment

informatics.kumc.edu
A Balanced Approach to R plug-ins for I2B2

- Kaplan Meier Web Client Plug-in
- A balanced approach to R plug-ins for I2B2
- I2B2 HIVE
- PM cell
- Kaplan Meier Web Client Plug-in
- I2B2 DW
- Biostatistics, R
- Patient privacy, institutional liability
- Python, SQL, HTML, JavaScript
- Cancer prevention, treatment
- Apache
- Biostatistics, R
- ABC
- XYZ
- Rpy libraries
- XYZ analysis.R
- KM analysis.R
Survival Comparison Between Cohorts

<table>
<thead>
<tr>
<th>Cohort 1</th>
<th>Cohort 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient Set:</strong></td>
<td>Resp. Cancer 35-64 [6-14-2013] [dconnolly] [PATIENTSET_24076]</td>
</tr>
<tr>
<td><strong>Sentinel event:</strong></td>
<td>Respiratory System [9,130 facts; 8,963 patients]</td>
</tr>
<tr>
<td><strong>Outcome:</strong></td>
<td>Deceased (Hospital Vital Status)</td>
</tr>
<tr>
<td></td>
<td>Death (Tumor Registry: NAACCR 1760)</td>
</tr>
<tr>
<td></td>
<td>Deceased (Social Security Death Index)</td>
</tr>
<tr>
<td></td>
<td>Other ...</td>
</tr>
<tr>
<td></td>
<td>Drop an outcome concept here</td>
</tr>
<tr>
<td><strong>Censor:</strong></td>
<td>Last Visii (any source)</td>
</tr>
<tr>
<td></td>
<td>Date of Last Contact (Tumor Registry: NAACCR 1750)</td>
</tr>
<tr>
<td></td>
<td>Other ...</td>
</tr>
<tr>
<td></td>
<td>Drop a censor event concept here</td>
</tr>
<tr>
<td><strong>Time Window:</strong></td>
<td>5 years</td>
</tr>
<tr>
<td><strong>Run Analysis</strong></td>
<td></td>
</tr>
</tbody>
</table>
## Survival Comparison Between Cohorts

The survival comparison between two cohorts was conducted using the Log-rank test. The results are as follows:

<table>
<thead>
<tr>
<th>Groups</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>440</td>
<td>197</td>
<td>210</td>
</tr>
<tr>
<td>2</td>
<td>259</td>
<td>125</td>
<td>111</td>
</tr>
</tbody>
</table>

The Chi-Square (Chisq) value is 2.676 on 1 degree of freedom, yielding a p-value of 0.102.

[Graph showing survival curves for two groups with the Kaplan-Meier method]
R Data Builder

Note: Access to RStudio Server is currently limited to members of the HERON study team.

Patient Set:
- Lung and Bronchus (8,911 cases, 6,842 patients)
- Never (371,137 cases, 141,196 patients)

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

Other Observations:
- Lung and Bronchus (8,911 cases, 6,842 patients)
- Never (371,137 cases, 141,196 patients)

Observations from Query:
- Lung and Bronchus (8,911 cases, 6,842 patients)

File:
- /home/dconnolly/heron/lung-never.smoked.Rda

Build R Data

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

- Name, Patient, n.obs
- Lung and Bronchus: 167, 171
- Never: 167, 1182
R Studio Server
Systems Architecture

- Identified data server
  - Staged source data
    - monthly refresh ETL
  - i2b2 compatible star schema
  - de-identification process
    - i2b2 compatible star schema
    - Application server
      - R scripts
        - plots, statistics
      - i2b2 Hive
        - rgate
      - Source System files (EMR dump, UHC CDB extract)
      - Investigator's client
        - One tab in browser
      - RStudio IDE web client
        - Another tab in browser
  - De-identified server
  - secure FTP/ETL

Source System files (EMR dump, UHC CDB extract)
Emerging Functionality: From Data Aggregation to Hospital Quality Preliminary Analysis

- R Data Builder plugin in i2b2 and integration with RStudio Server
  - (http://www.rstudio.com/ide/docs/server/getting_started)

- Test Case: **Antibiotic Administration for Septic patients in the Emergency Room**
  - Past publication to bring in flowsheet data an important foundation
  - University HealthSystem Consortium CDB “gold” standard for KU Hospital
  - What can you solve in i2b2 “same financial encounter” versus send to R?
Repurposing i2b2 Clinical Research Infrastructure for Inpatient Quality Improvement

- I2b2 originally ambulatory or population/genomics focused
- Is i2b2 version 1.6 with same financial encounter and modifiers now useful for inpatient research?

- Goal: understand medication timing and antibiotic selection
- Suspect vancomycin preferred
- Validate HERON medications
  - Especially administration timing
R Data Builder Plugin and RStudio Server

Web based for user. Just another tab in the browser.

All data stays on the server so there’s no data release and risk of re-identification due to a lost file.

i2b2 Plugin invokes a program that creates a Rda file in their directory on the server.
3513 patients had a UHC-defined sepsicemia diagnosis

2912 patients were an Emergency Admission

2861 patients age were 18 years or older

2722 patients had an exposure to an Antibiotic in the encounter

1839 had ED Triage documentation during the encounter

1836 had the Sepsis Screen Used during the encounter

1223 had 2 SIRS criteria, organ dysfunction and suspicion/treatment of infection

717 MD notified

Average time to sepsis screening 2.9 hours, median 49 minutes

Cohorts above line defined with i2b2

Cohorts below line further refined with R

Note: 28 patients who lacked an ED departure time were excluded from further analysis

1244 patients had 1st antibiotic admin within 24 hours (1474 encounters)

Average time in ED is 7.9 hours, median 7.1

261 had 1st antibiotic admin before sepsis screening (277 encounters)

1040 had 1st antibiotic admin after sepsis screening (1197 encounters)

993 had 1st antibiotic admin given in ED (1140 encounters)

316 had 1st antibiotic admin not in ED (334 encounters)

Average time spent in ED is 8.7 hours, median 7.6

Average time spent in ED is 6.7 hours, median 6.6
Density Plots: Time from Arrival to First Antibiotic

Broad Spectrum versus Vancomycin

Lag when given outside Emergency Room

Lag in Broad Spectrum after Vancomycin

Administration relative to RN Sepsis Screen

When
- in.ed
- not.in

Admin
- before
- after
• REDCap registries into i2b2 allows intuitive exploration
  – Researchers may need less abstraction as data is extracted from the EMR.
• i2b2 into REDCap: inherit security model, graphical/export tools
1st Annual HERON Fishing Tournament: HERON training workshop

• August 1st:
  o 1:00-5:00 PM Classroom style training
  o Convene for social gathering

• August 2nd:
  o 8:00-0:00 AM Hands-on training on attendee topics
  o 10:15-11:30 Discussion

Invited: KUMC/Frontiers researchers and regional informaticians

frontiersresearch.org  informatics.kumc.edu