RGC Database

Isaac Hodes, Sarah Kerns, Barry Rosenstein

RGC Conference, Montpellier France, June 2015

Hi!

I’m a software engineer.

Background

- Software Engineer with mathematics background
- Hammer Lab out of the Icahn School of Medicine at Mount Sinai
  - Our mission is to improve the state and standard of care for patients by making high-quality (tested, resilient, well-designed) software for doctors and researchers
  - All our work is out in the open, free, & open source on GitHub
  - Group of engineers with, mostly, backgrounds in computer science & math, from e.g. Google, Facebook, Foursquare, Airbus, etc.
  - We want to use data to make healthcare a better place
- Some other projects
  - We work on the computational pipeline for our cancer vaccine
  - Also provide a Hadoop compute cluster for Multiscale Genomics
  - Variant calling analysis framework called CycleDash
  - Exploring variant calling using new distributed technologies

What am I doing here?

Well…

RGC Database

On the RGC NCI Website there are five stated goals

1. Fostering international collaborative research projects in radiogenomics through sharing of biospecimens and data;
2. Developing guidelines to improve the standardization of radiogenomics research;
3. Providing a framework for the efficient conduct and publication of original data meta-analyses of relevant studies;
4. Providing a forum and framework for discussion, development and pursuit of new research directions; and
5. Supporting the development of early career researchers.

RGC Database

On the RGC NCI Website there are five stated goals:

1. *Fostering international collaborative research projects in radiogenomics through sharing of biospecimens and* **data***;*
2. Developing guidelines to improve the standardization of radiogenomics research;
3. Providing a framework for the efficient conduct and publication of original data meta-analyses of relevant studies;
4. Providing a forum and framework for discussion, development and pursuit of new research directions; and
5. Supporting the development of early career researchers.

Why is this hard?

- **It's difficult to share data.**
- **Researchers spend much of their time wrangling study data that is**
  - in different formats/structures
  - time-consuming to gather
  - using different measurement scales
  - measuring different endpoints
  - undocumented

Consolidating Consortium Data

- Site data is measuring similar things
  - Patient demographics — age, race
  - Comorbidities — smoker, diabetic
  - Treatment information — hormone therapy, EBRT
  - Toxicity measurements — hematuria, GI ulcer
  - Genotype data

Consolidating Consortium Data

- In reality there is modest overlap between
  - Measurements/endpoints
    - ⅖ studies measured hematuria
    - ⅗ studies measure frequency of urination, but with 3 different scales
  - Scales
    - IPSS
    - LENT-SOM
    - CTCAEv2 (and v3, v4)
    - RTOG/EORTC
    - SHIM
    - Custom scales (binary, yes/no, other)
- This leads to data that can't be (easily) harmonized.
- In addition, there's the challenge of measurements over time
  - How can we easily ask questions of data over time?
  - How do we deal with the varying time periods between studies?

And so, the RGC Database

RGC Database

The goal is to…

1. Process data from many studies (& automate this processing)
2. Expose that data to other consortium members (with permission)
3. Enable querying and inspection of the data
    a. and allowing relevant subsets to be selected
4. Allow export of that data to common formats (CSV, XLS, Stata, R, etc.)
5. Update and add new studies & cancer types (starting with prostate)

RGC Database

- Solution?
  - A comprehensive ETL (extract, transform, and load) pipeline for transforming and verifying data before loading it into the database
  - A web-based secure, modern web application using proven technology
    - Uses a project from the Children's Hospital of Philadelphia called "Harvest"
    - Modern technologies and best practices: Python frameworks and Postgres database hosted in the cloud
    - Access controls to ensure only the data a user has permission to view is available

1 — Process

Process data from many studies (& automate this processing)

- This is primarily a matter of wrangling data from the various formats studies use into one standard format the database can understand
- Users can drag their formatted spreadsheet into the browser and have it automatically verified and uploaded (once approved) to the database

2 & 3 — Viewing & Querying

Enable querying and inspection of the data

- This is where the power of Harvest and a uniform data format help us
- Simple queries include
  - “patients who got radiotherapy before the age of 60 and are not smokers”
  - “patients who reported hematuria and have diabetes”

4 — Export

Allow export of that data to common formats (CSV, XLS, Stata, R, etc.)

- The results of any query can be exported to any common format
- On this data the researcher may then locally run their analysis

5—Update

Update and add new studies & cancer types (starting with prostate)

- New and updated studies can be easily added to the database by researchers

Demo Time!

Fields

Toxicity Measurements −

GENITOURINARY

- Dysuria
- Hematuria
- IPSS Total
- Quality of Life
- Frequency (GU)
- Incomplete emptying (GU)
- Incontinence (GU)
- Nocturia (GU)
- Intermittency (GU)
- Cystitis
- Stricture (GU)
- Quality of Life
- Retention (GU)
- Straining (GU)
- Urgency (GU)
- Weak Stream (GU)

GASTROINTESTINAL

- Diarrhea
- Proctitis
- Rectal bleeding
- Anal Incontinence
- Fistula (GI)
- Hemorrhage (Gi)
- Perforation (GI)
- Stricture (GI)
- Ulcer (GI)

SEXUAL FUNCTION

- SHIM Confidence
- SHIM Total
- Potency

Treatment −

- Date Of Turp
- Hormones
- Hormones Months Before
- Hormones Months After
- Implant Date
- Implant Type
- Isotope
- Ebrt Date Start
- Ebrt Dose Cgy
- Ebrt Fraction
- Ebrt Number

Comorbidities −

- Smoker
- Alcohol
- Hypertension
- Diabetes
- Heart Disease

Demographic −

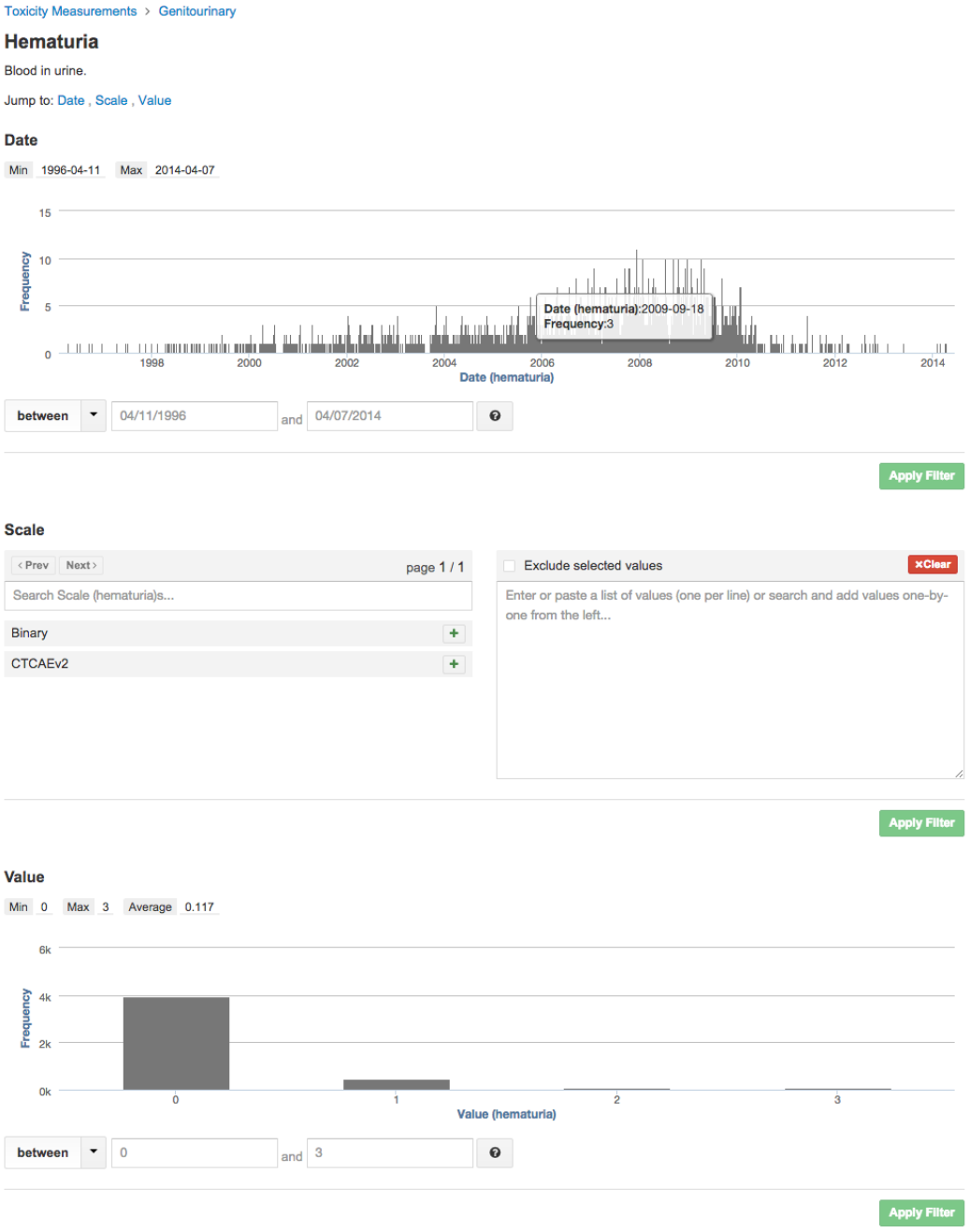
- Internal Id
- Age At Rx
- Race
- Gleason Score
- Initial Psa
- T Stage
- N Stage
- M Stage
- Pre Hormones Volume
- Dataset Name

Querying

Dates

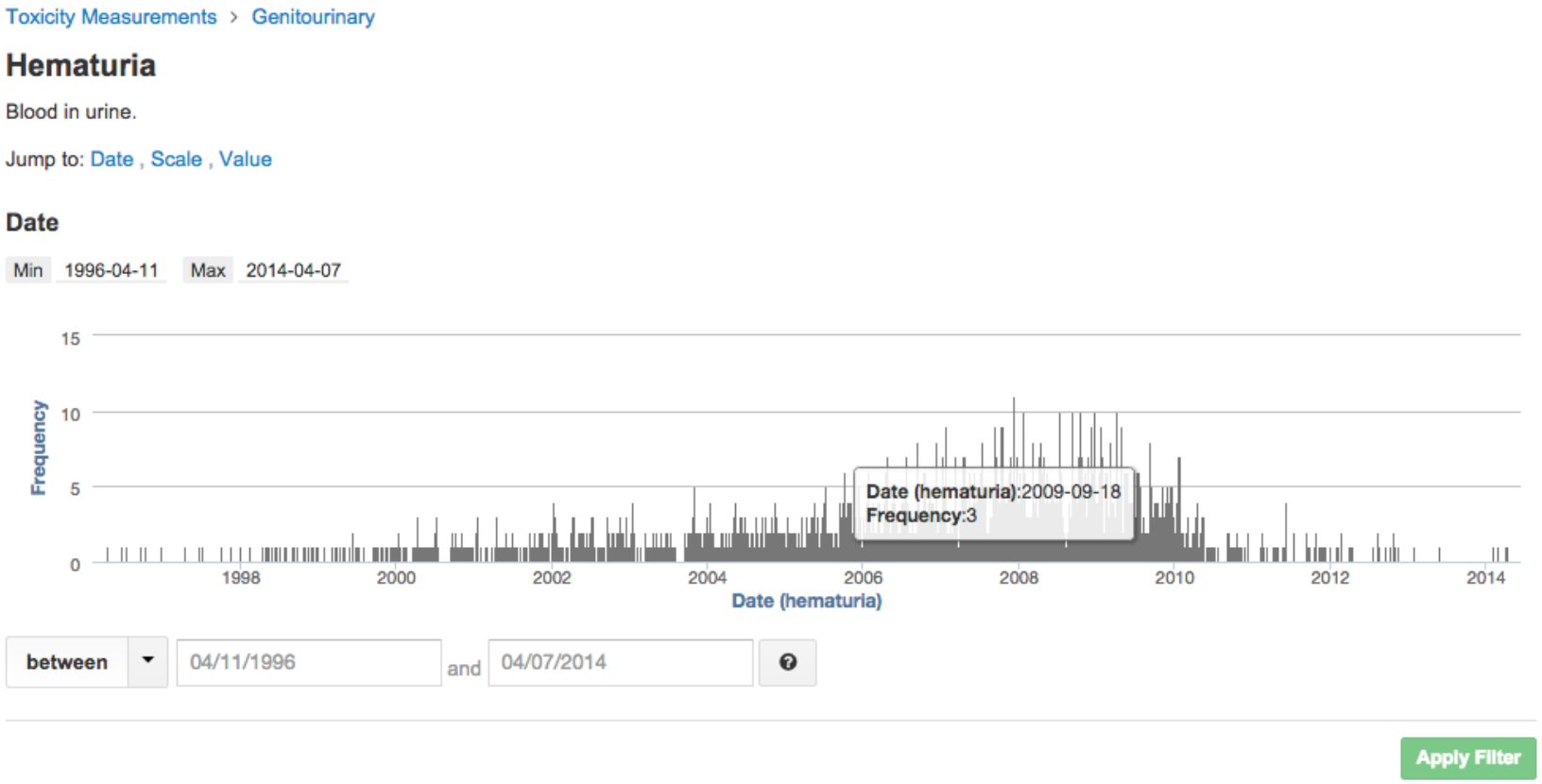
Scales

Values



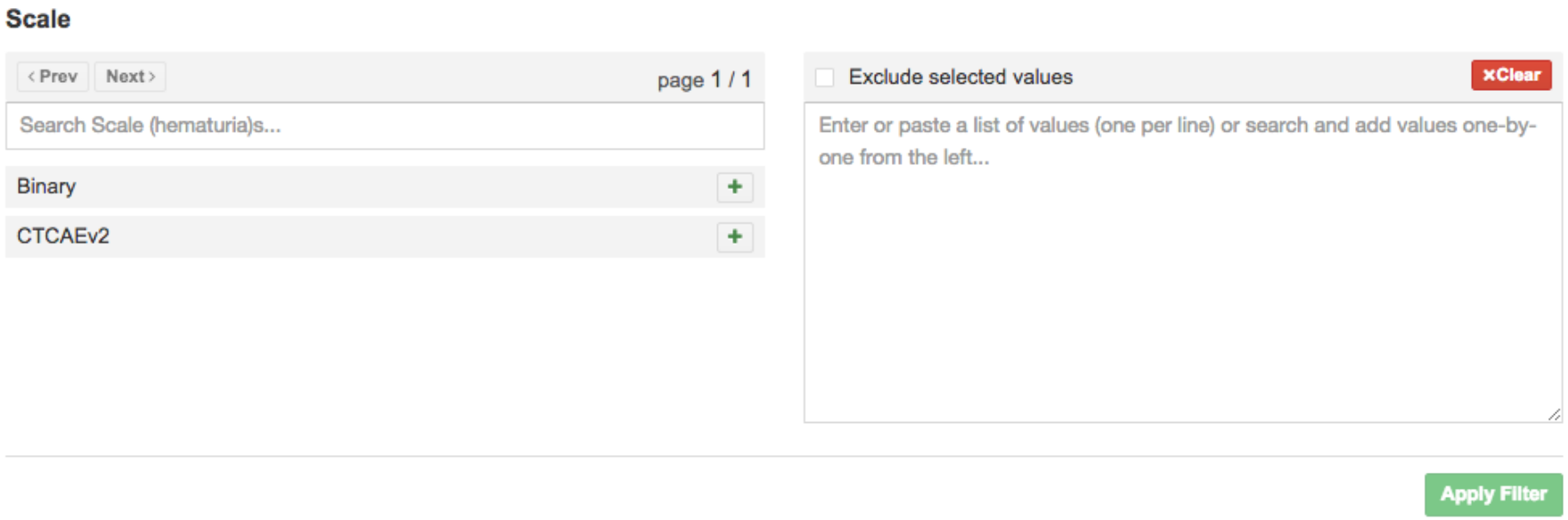
Querying

Select measurements between certain dates



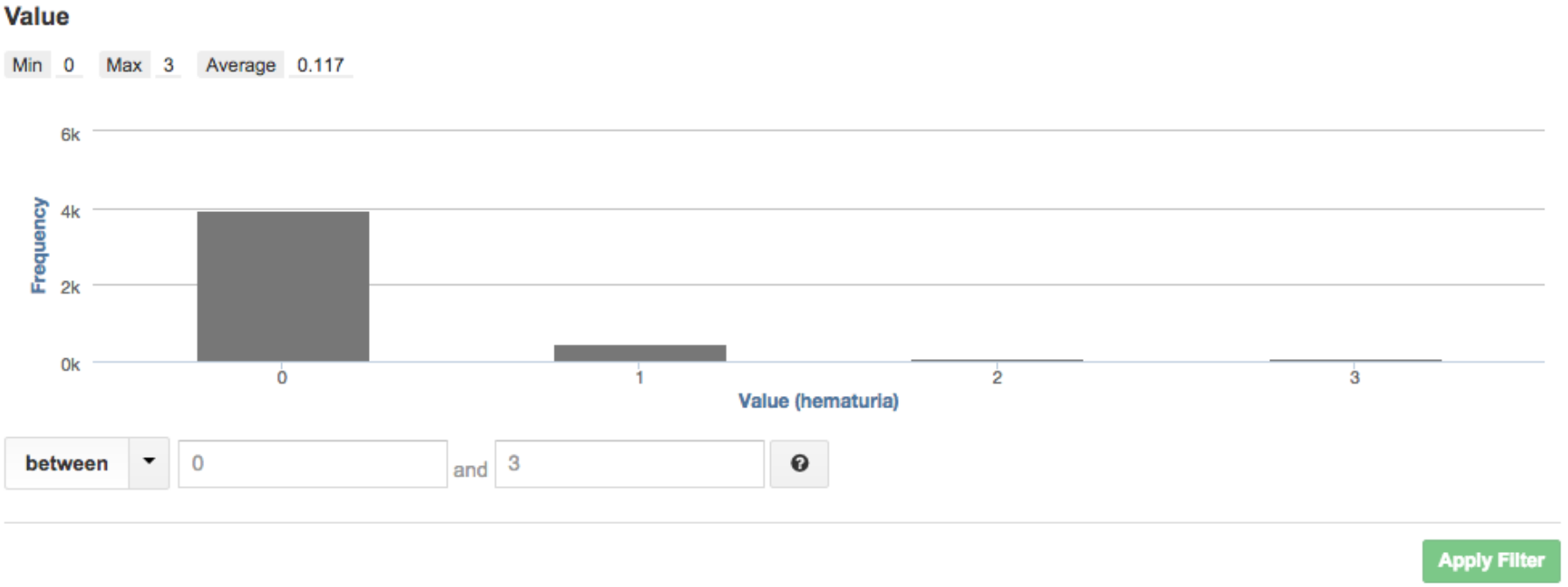
Querying

Select measurements with a given scale



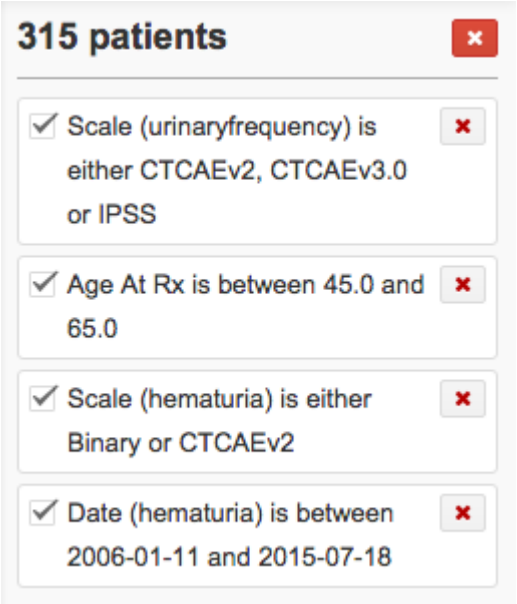
Querying

Select measurements with a range of values



Querying

As you edit your query, the number of patients matching your filters changes



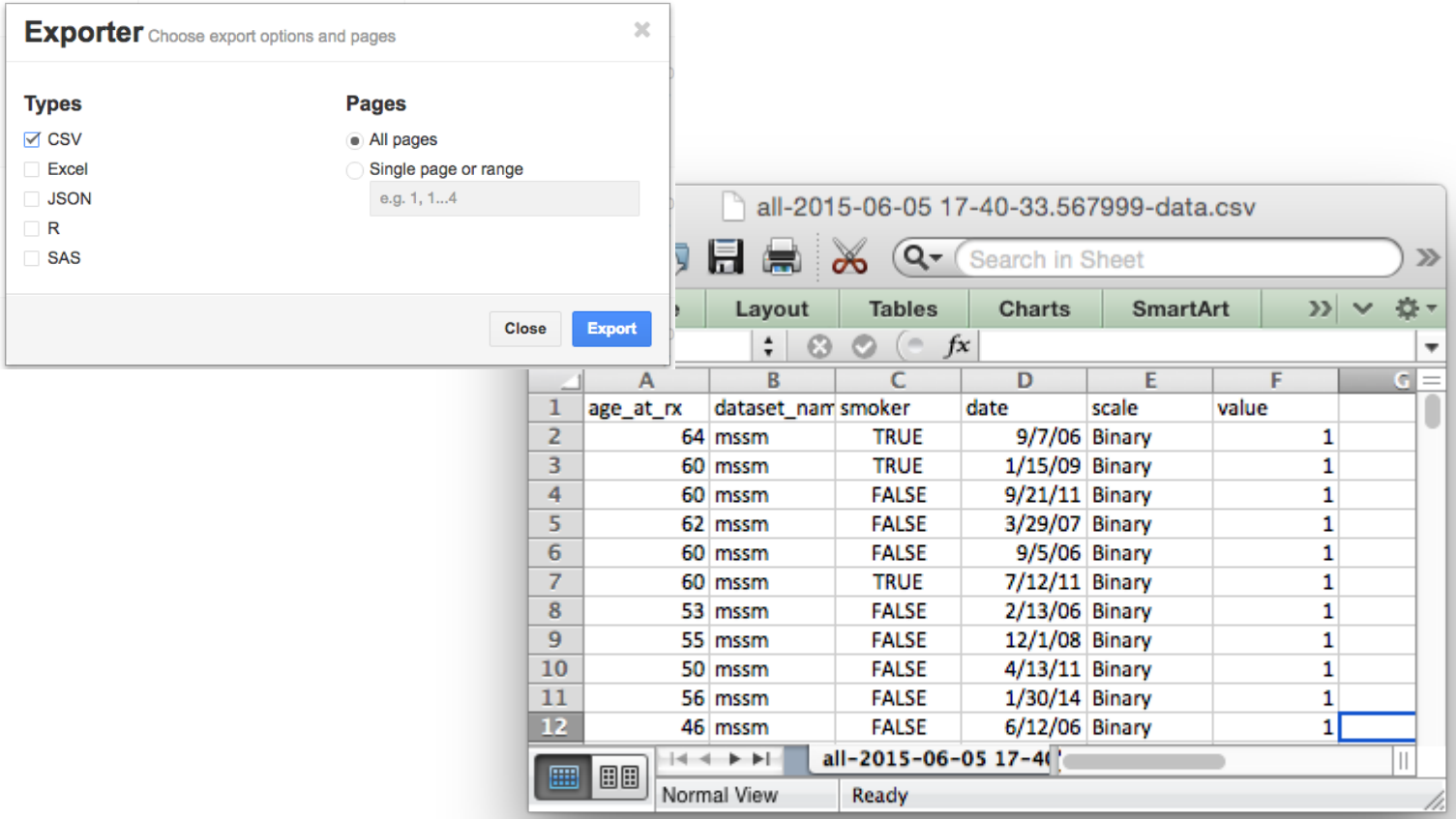
Results

143 records

« ‹ › » page 1 / 8

Age At Rx	Dataset Name	Smoker	Hematuria
56	mssm	False	**Date** 2014-01-30 **Scale** Binary **Value** 1
59	mssm	False	**Date** 2009-06-15 **Scale** Binary **Value** 1
65	mssm	False	**Date** 2011-06-01 **Scale** Binary **Value** 1
55	mssm	False	**Date** 2008-10-09

Exporting



	A	B	C	D	E	F
1	age_at_rx	dataset_nam	smoker	date	scale	value
2	64	mssm	TRUE	9/7/06	Binary	1
3	60	mssm	TRUE	1/15/09	Binary	1
4	60	mssm	FALSE	9/21/11	Binary	1
5	62	mssm	FALSE	3/29/07	Binary	1
6	60	mssm	FALSE	9/5/06	Binary	1
7	60	mssm	TRUE	7/12/11	Binary	1
8	53	mssm	FALSE	2/13/06	Binary	1
9	55	mssm	FALSE	12/1/08	Binary	1
10	50	mssm	FALSE	4/13/11	Binary	1
11	56	mssm	FALSE	1/30/14	Binary	1
12	46	mssm	FALSE	6/12/06	Binary	1

Data Modeling

Consolidating Consortium Data

- But how do we model the data?
- A few options…
  - Harmonize the data and try to combine data
  - Leave it be; a collection of unrelated spreadsheets and databases
  - Somewhere in between
- Clearly, the latter option was what we chose

The New Schema

- Data split into a few logical components with a consistent schema
  - Patient demographic &tc data
    - Age at Rx, race, cancer type
    - Comorbidities
  - Treatment data
    - Hormones, external beam radiation, implanted radiation
  - Toxicity data
    - Time-series data of e.g. incontinence, rectal bleeding, hematuria
  - Genotype data

The New Schema

- The result is a flexible, *sparse*, standard format
- With this we can
  - Display data in a readable and reasonable fashion
  - Export data into easily-manipulable formats
  - Import unknown additional data types with no changes required
  - Handle time-series data taken over any period of time
- This also allows us to easily incorporate new data from new studies without losing any of the above advantages

Wrapping up

What's Next?

- Incorporate genotype data
- Interface to upload additional data in standard formats
- Work out permissioning and authorization
- Continue improving the web interface

Initial Studies & Thanks

MSSM — PIs: Barry Rosenstein, PhD and Harry Ostrer, MD
Institution: Icahn School of Medicine at Mount Sinai, New York, USA

RADIOGEN — PI: Ana Vega, PhD
Institution: Universidade de Santiago de Compostela, Santiago de Compostela, Spain

CCI — PIs: Matthew Parliament, MD and Nawaid Usmani, MD
Institution: University of Alberta Cross Cancer Institute, Edmonton, Canada

NTMC — PI: Shiro Saito, MD
Institution: National Hospital Organization Tokyo Medical Center, Tokyo, Japan

NIRS — PI: Takashi Imai, MD
Institution: National Institute of Radiological Sciences, Chiba, Japan

RAPPER — PIs: Catharine West, PhD and Neil Burnet, MD
Institution: multi-site throughout the UK and Europe
(Includes RT01, CHHiP, RADICALS, Dose Escalation, PelvicIMRT, PIVOTAL, STAMPEDE)

Appendix

Patient data format

patient_id	internal_id	study_name	race	age_at_rx	etc…
341	89	nirs	japanese	67	…
1009	D067E1	radiogean	caucasian	75	…
197	009E11	mssm	hispanic	71	…

Toxicity data format

Toxicity Table (e.g. Hematuria)

Patient ID	Scale	Date	Value
417	CTCAEv2	2015-04-03	4
413	Binary	2013-09-23	1
102	CTCAEv3	2013-10-11	2
1078	CTCAEv2	1998-12-12	0