# Teaching OHDSI in a University Course: Lessons Learned at Georgia Tech

OHDSI Community Presentation

10/29/2019

Jon Duke, MD

# GT Masters in Computer Science

- Georgia Tech has the largest Computer Science graduate program in the US
- In 2014, GT started the Online Master's in Computer Science (OMSCS)
  - – OMSCS degree costs $7K vs ~$40K on-campus

| Applications TO DATE | Enrollment AS OF SPR2019 | # of Countries REPRESENTED | OMSCS News Articles |
|---|---|---|---|
| 26k | 8664 | 114 | 1200+ |

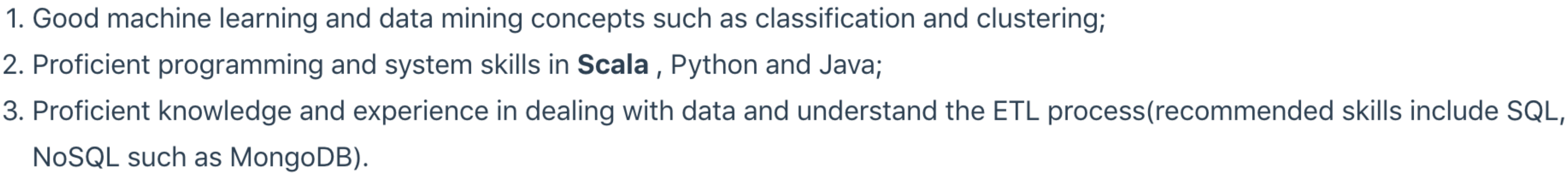
# CS6440: Intro to Health Informatics

- Broad introduction to EHRs, the US healthcare system, healthcare quality, healthcare data and vocabularies
  - Started by Dr. Mark Braunstein in 2012
  - Taught in OMSCS and on-campus
  - Strong focus on FHIR and Interoperability
- Student majors 85% Comp Sci and remainder including biomedical engineering, HCI, bioinformatics, industrial engineering

# OHDSI in CS6440

- I took over the class in 2018
  - Decided to add an OHDSI block for Fall 2019 semester
- NB: GT has a more ‘hardcore’ health data analytics course taught by Dr. Jimeng Sun
  - [Big Data for Healthcare]()

CSE6250 Prerequisites



1. Good machine learning and data mining concepts such as classification and clustering;
2. Proficient programming and system skills in **Scala** , Python and Java;
3. Proficient knowledge and experience in dealing with data and understand the ETL process(recommended skills include SQL, NoSQL such as MongoDB).

# CS6440 Fall 2019

- People
  - 386 students
  - 14 TAs
  - Me
- Course Educational Infrastructure
  - Canvas (assignments, submissions)
  - Udacity (lectures)
  - Youtube (lectures)
  - Piazza (forum)
  - Slack

# Goals of the OHDSI Block

- Learn the **kinds of questions** people ask using observational data (the OHDSI trinity)
- Get **hands-on experience using the OHDSI framework** to answer a question of your own
- **Get excited about the possibilities** of how health data can be used in FHIR application development (second part of the course)

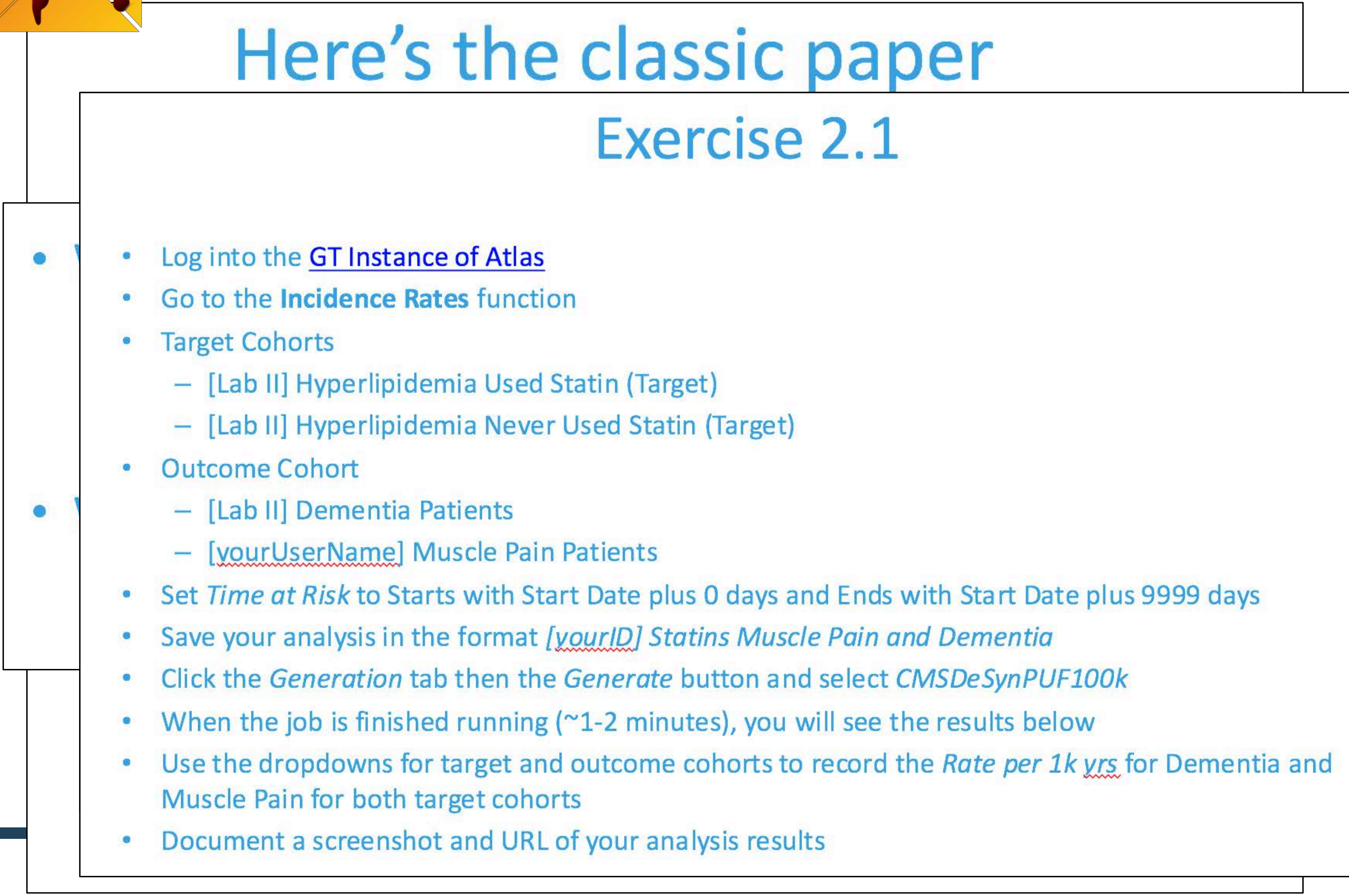
# Non-Goals of the OHDSI Block

- Become an expert in medicine / epi / stats / clinical research
- OHDSI best practices, conventions, ETL design, etc

# Components of the Analytics Block

- Data Standards lectures and activities
- OHDSI Labs (slides, videos, exercises)
  - Intro
  - Lab I: Concept Set Design
  - Lab II: Cohort Design and Characterization
  - Lab III: Incidence Rates and Estimation Study
- Individual Health Analytics Project
  - Proposal, Design, Execution, Report

# Examples from Lab

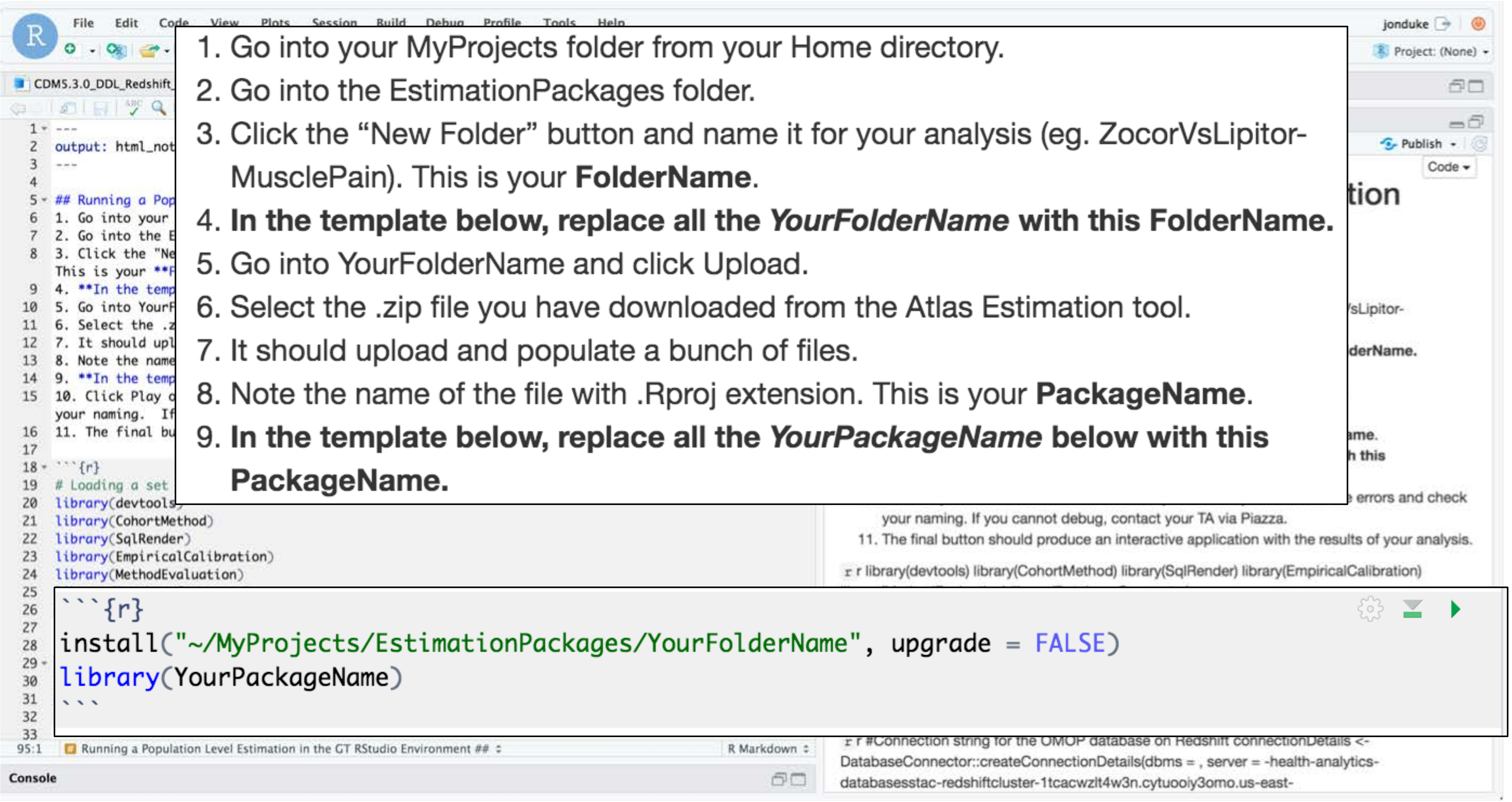


## Here’s the classic paper

## Exercise 2.1

- Log into the GT Instance of Atlas
- Go to the **Incidence Rates** function
- Target Cohorts
  - [Lab II] Hyperlipidemia Used Statin (Target)
  - [Lab II] Hyperlipidemia Never Used Statin (Target)
- Outcome Cohort
  - [Lab II] Dementia Patients
  - [yourUserName] Muscle Pain Patients
- Set *Time at Risk* to Starts with Start Date plus 0 days and Ends with Start Date plus 9999 days
- Save your analysis in the format *[yourID] Statins Muscle Pain and Dementia*
- Click the *Generation* tab then the *Generate* button and select *CMSDeSynPUF100k*
- When the job is finished running (~1-2 minutes), you will see the results below
- Use the dropdowns for target and outcome cohorts to record the *Rate per 1k yrs* for Dementia and Muscle Pain for both target cohorts
- Document a screenshot and URL of your analysis results

# PLE Markdown Template for our Analytics Environment



1. Go into your MyProjects folder from your Home directory.
2. Go into the EstimationPackages folder.
3. Click the “New Folder” button and name it for your analysis (eg. ZocorVsLipitor-MusclePain). This is your **FolderName**.
4. **In the template below, replace all the *YourFolderName* with this FolderName.**
5. Go into YourFolderName and click Upload.
6. Select the .zip file you have downloaded from the Atlas Estimation tool.
7. It should upload and populate a bunch of files.
8. Note the name of the file with .Rproj extension. This is your **PackageName**.
9. **In the template below, replace all the *YourPackageName* below with this PackageName.**

```
# Loading a set
library(devtools)
library(CohortMethod)
library(SqlRender)
library(EmpiricalCalibration)
library(MethodEvaluation)
```

```{r}
install("~/MyProjects/EstimationPackages/YourFolderName", upgrade = FALSE)
library(YourPackageName)
```

95:1 Running a Population Level Estimation in the GT RStudio Environment ##

errors and check your naming. If you cannot debug, contact your TA via Piazza.

11. The final button should produce an interactive application with the results of your analysis.

r library(devtools) library(CohortMethod) library(SqlRender) library(EmpiricalCalibration)

r #Connection string for the OMOP database on Redshift connectionDetails <- DatabaseConnector::createConnectionDetails(dbms = , server = -health-analytics-databasesstac-redshiftcluster-1tcacwzlt4w3n.cytuooiy3omo.us-east-

# Example Submission

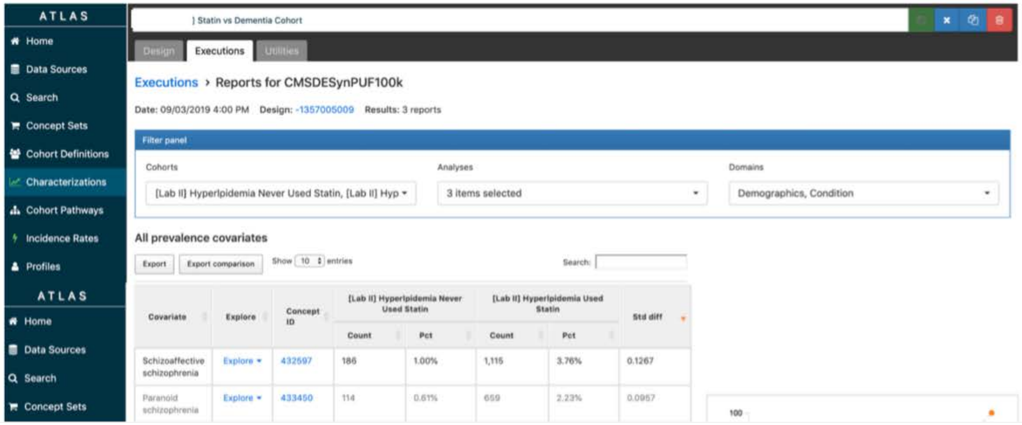
09/03/2019

Exercise 2.2:

What is the age group with the biggest difference (Std Diff) between the Statin Users and Non-Statin Users? 50-54

What condition shows the greatest difference (Std Diff) between Statin Users and Non-Statin Users? Schizoaffective schizophrenia

http://gt-health-analytics.us-east-1.elasticbeanstalk.com/#/cc/characterizations/46/results/867



Executions › Reports for CMSDESynPUF100k

Date: 09/03/2019 4:00 PM Design: -1357005009 Results: 3 reports

Cohorts: [Lab II] Hyperlpidemia Never Used Statin, [Lab II] Hyp

Analyses: 3 items selected

Domains: Demographics, Condition

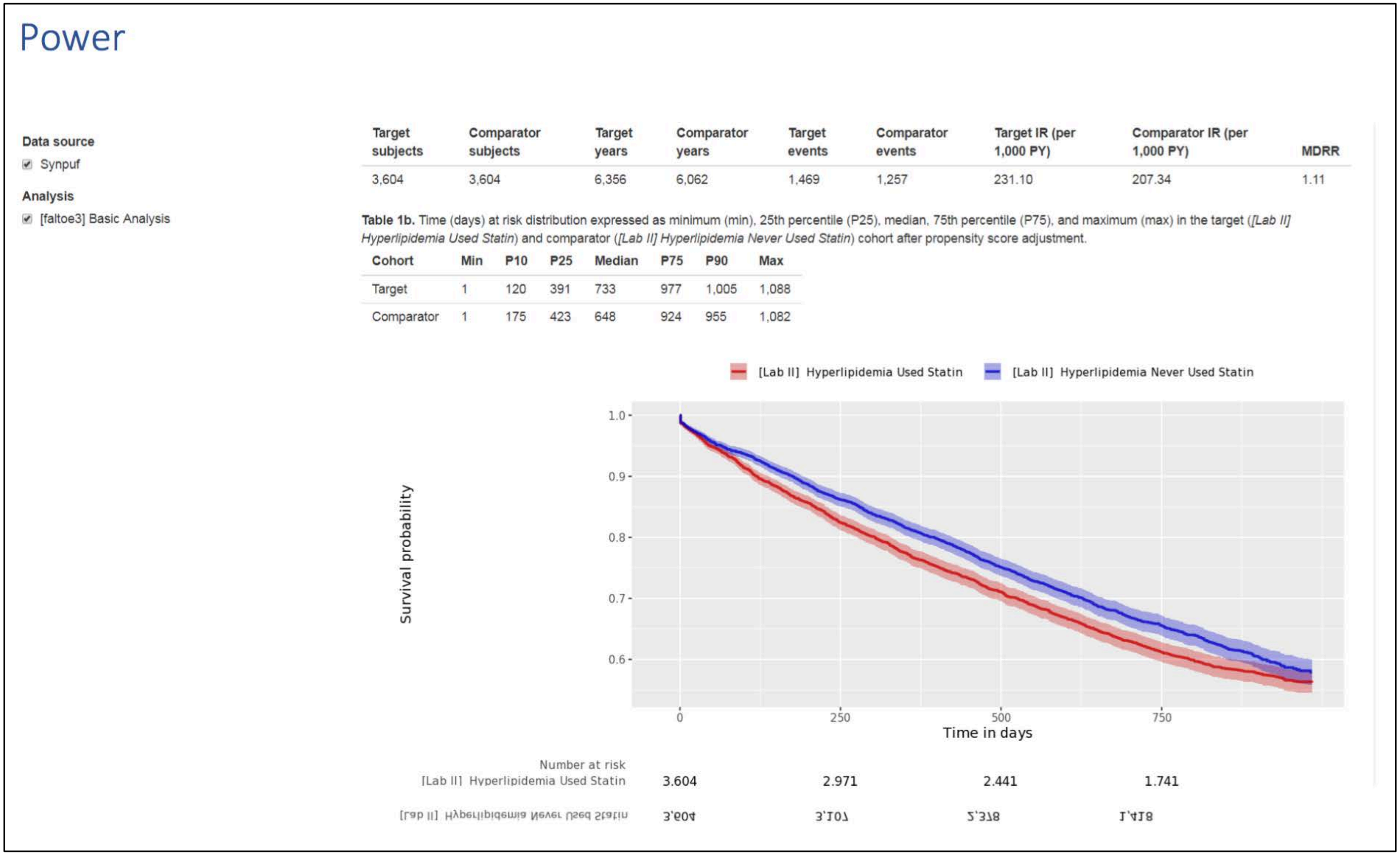
All prevalence covariates

Show 10 entries

Search:

| Covariate | Explore | Concept ID | [Lab II] Hyperlpidemia Never Used Statin | | [Lab II] Hyperlpidemia Used Statin | | Std diff |
|---|---|---|---|---|---|---|---|
| | | | Count | Pct | Count | Pct | |
| Schizoaffective schizophrenia | Explore ▾ | 432597 | 186 | 1.00% | 1,115 | 3.76% | 0.1267 |
| Paranoid schizophrenia | Explore ▾ | 433450 | 114 | 0.61% | 659 | 2.23% | 0.0957 |

# Example Submission



## Power

**Data source**

☑ Synpuf

**Analysis**

☑ [faltoe3] Basic Analysis

| Target subjects | Comparator subjects | Target years | Comparator years | Target events | Comparator events | Target IR (per 1,000 PY) | Comparator IR (per 1,000 PY) | MDRR |
|---|---|---|---|---|---|---|---|---|
| 3,604 | 3,604 | 6,356 | 6,062 | 1,469 | 1,257 | 231.10 | 207.34 | 1.11 |

**Table 1b.** Time (days) at risk distribution expressed as minimum (min), 25th percentile (P25), median, 75th percentile (P75), and maximum (max) in the target (*[Lab II] Hyperlipidemia Used Statin*) and comparator (*[Lab II] Hyperlipidemia Never Used Statin*) cohort after propensity score adjustment.

| Cohort | Min | P10 | P25 | Median | P75 | P90 | Max |
|---|---|---|---|---|---|---|---|
| Target | 1 | 120 | 391 | 733 | 977 | 1,005 | 1,088 |
| Comparator | 1 | 175 | 423 | 648 | 924 | 955 | 1,082 |

🟥 [Lab II] Hyperlipidemia Used Statin
🟦 [Lab II] Hyperlipidemia Never Used Statin
Survival probability
1.0
0.9
0.8
0.7
0.6
0
250
500
750
Time in days

| Number at risk | | | | |
|---|---|---|---|---|
| [Lab II] Hyperlipidemia Used Statin | 3.604 | 2.971 | 2.441 | 1.741 |
| [Lab II] Hyperlipidemia Never Used Statin | 3.604 | 3.107 | 2.318 | 1.418 |

# Individual Health Analytics Project

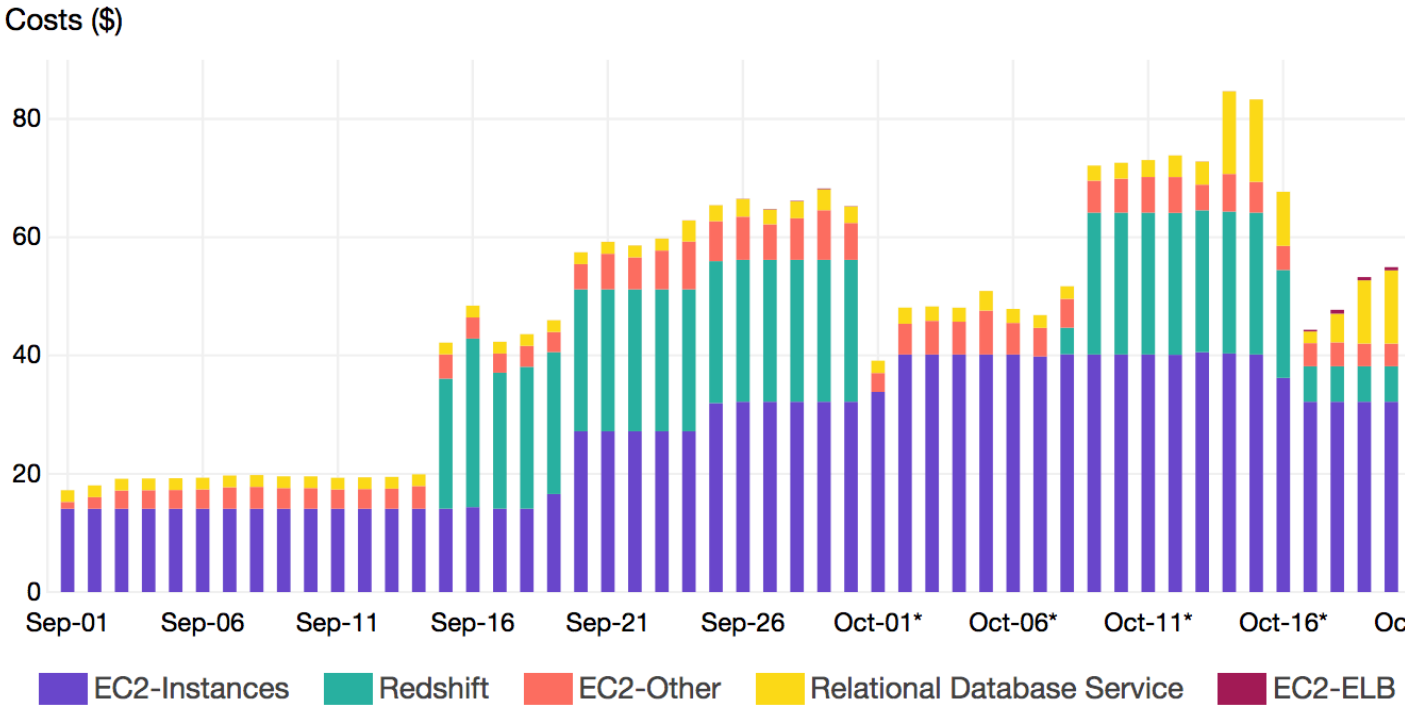
- Propose a T vs C for outcome O question appropriate for SynPUF dataset
- Create concept sets and cohorts
- Perform Atlas Characterization and Incidence
- Generate Estimation Study and run in R
- Write a Report

# Our OHDSI Stack: OHDSI on AWS

- OMOP CDM
  - SynPUF 100k/2.3M
  - Redshift dc2.large x 2 nodes (later 4 nodes)
- Atlas
  - Elastic Beanstalk
    - t3.medium x 2-4 nodes (later t3.2xlarge x 2 nodes)
  - OHDSI Schema DB
    - RDS Aurora Postgres db.t3.medium (later r5.4xlarge)
- Rstudio
  - R5.4xlarge
  - 500GB (later 750GB)

# Costs

- Initial costs ~$20/day
- Project peaks $50-75/day

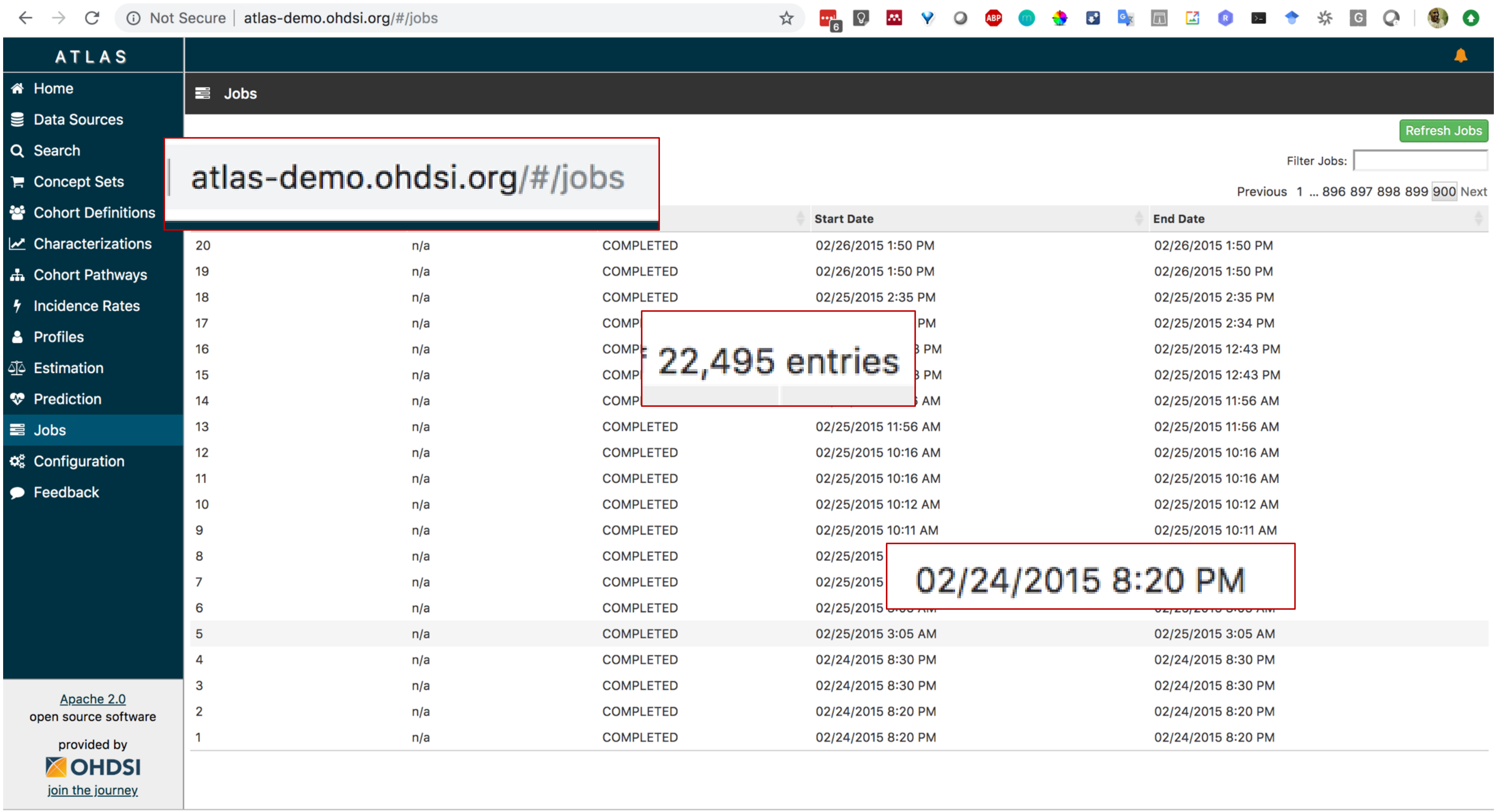


# Authentication

- We used Atlas security (Shiro)
- Each student was assigned a username / pw
- Does not hide other students’ work, so all is visible to all
- But does let us track who did what when
- OHDSIonAWS sets up automatically same credentials for Atlas and RStudio

# So how did it go?

# For Reference
# Atlas Jobs on ohdsi.org



atlas-demo.ohdsi.org/#/jobs

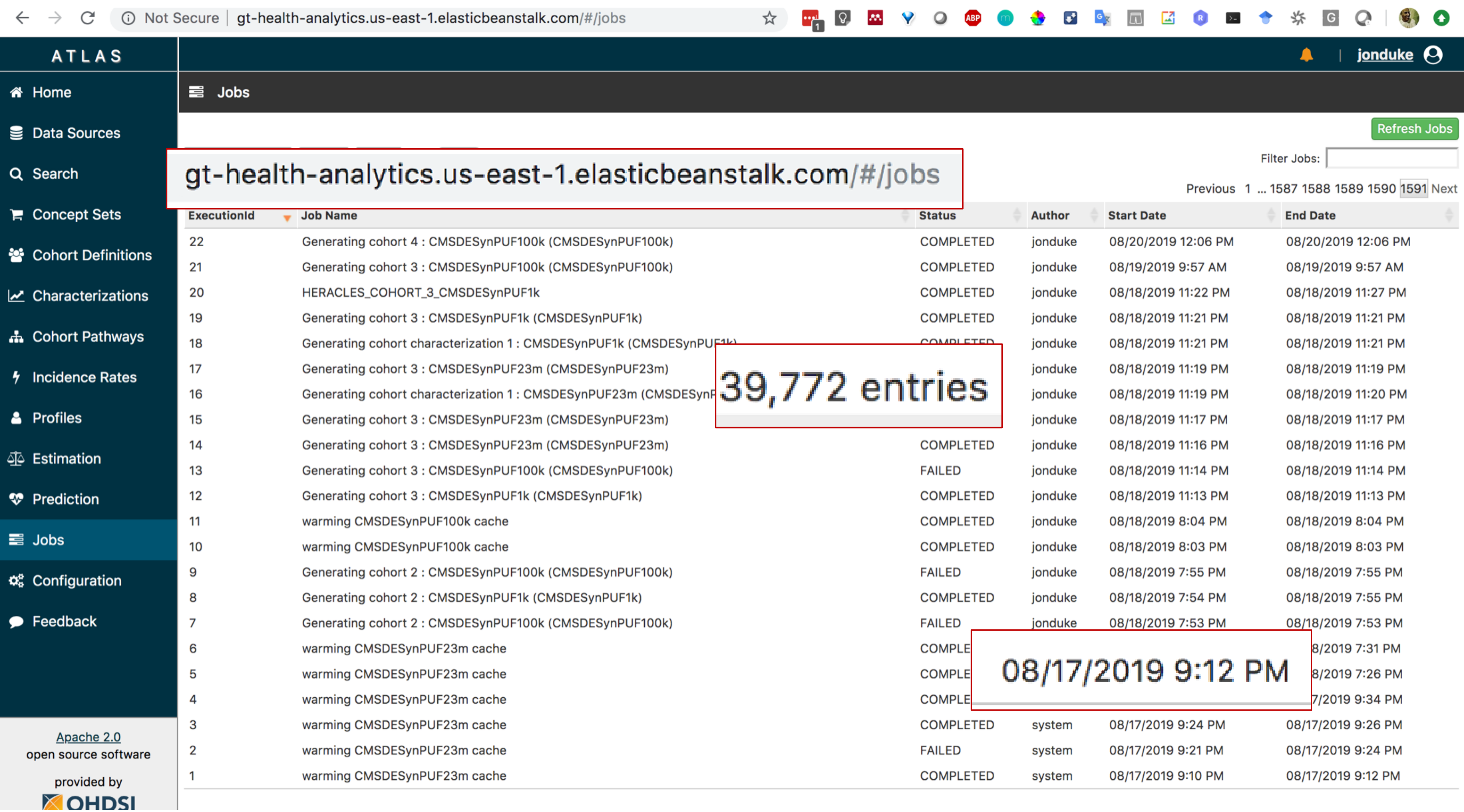
22,495 entries

02/24/2015 8:20 PM

| | | | Start Date | End Date |
|---|---|---|---|---|
| 20 | n/a | COMPLETED | 02/26/2015 1:50 PM | 02/26/2015 1:50 PM |
| 19 | n/a | COMPLETED | 02/26/2015 1:50 PM | 02/26/2015 1:50 PM |
| 18 | n/a | COMPLETED | 02/25/2015 2:35 PM | 02/25/2015 2:35 PM |
| 17 | n/a | COMP | PM | 02/25/2015 2:34 PM |
| 16 | n/a | COMP | 3 PM | 02/25/2015 12:43 PM |
| 15 | n/a | COMP | 3 PM | 02/25/2015 12:43 PM |
| 14 | n/a | COMP | 5 AM | 02/25/2015 11:56 AM |
| 13 | n/a | COMPLETED | 02/25/2015 11:56 AM | 02/25/2015 11:56 AM |
| 12 | n/a | COMPLETED | 02/25/2015 10:16 AM | 02/25/2015 10:16 AM |
| 11 | n/a | COMPLETED | 02/25/2015 10:16 AM | 02/25/2015 10:16 AM |
| 10 | n/a | COMPLETED | 02/25/2015 10:12 AM | 02/25/2015 10:12 AM |
| 9 | n/a | COMPLETED | 02/25/2015 10:11 AM | 02/25/2015 10:11 AM |
| 8 | n/a | COMPLETED | 02/25/2015 | |
| 7 | n/a | COMPLETED | 02/25/2015 | |
| 6 | n/a | COMPLETED | 02/25/2015 | |
| 5 | n/a | COMPLETED | 02/25/2015 3:05 AM | 02/25/2015 3:05 AM |
| 4 | n/a | COMPLETED | 02/24/2015 8:30 PM | 02/24/2015 8:30 PM |
| 3 | n/a | COMPLETED | 02/24/2015 8:30 PM | 02/24/2015 8:30 PM |
| 2 | n/a | COMPLETED | 02/24/2015 8:20 PM | 02/24/2015 8:20 PM |
| 1 | n/a | COMPLETED | 02/24/2015 8:20 PM | 02/24/2015 8:20 PM |

As of 10/14/2019

# Atlas Jobs on GT OHDSI



gt-health-analytics.us-east-1.elasticbeanstalk.com/#/jobs

39,772 entries

08/17/2019 9:12 PM

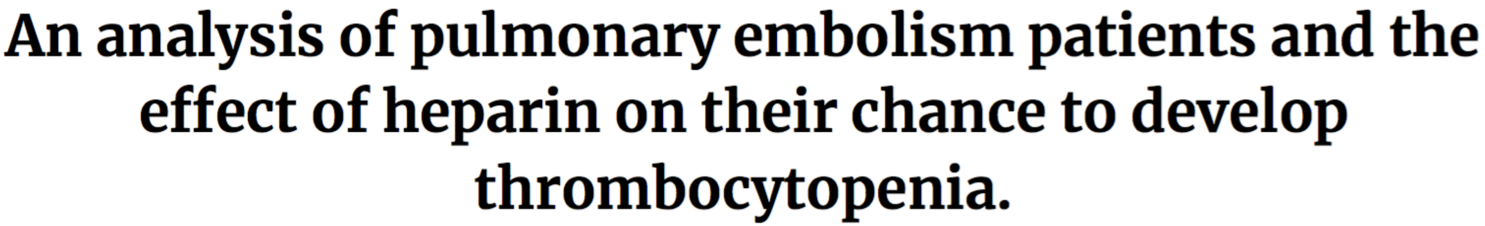
| ExecutionId | Job Name | Status | Author | Start Date | End Date |
|---|---|---|---|---|---|
| 22 | Generating cohort 4 : CMSDESynPUF100k (CMSDESynPUF100k) | COMPLETED | jonduke | 08/20/2019 12:06 PM | 08/20/2019 12:06 PM |
| 21 | Generating cohort 3 : CMSDESynPUF100k (CMSDESynPUF100k) | COMPLETED | jonduke | 08/19/2019 9:57 AM | 08/19/2019 9:57 AM |
| 20 | HERACLES_COHORT_3_CMSDESynPUF1k | COMPLETED | jonduke | 08/18/2019 11:22 PM | 08/18/2019 11:27 PM |
| 19 | Generating cohort 3 : CMSDESynPUF1k (CMSDESynPUF1k) | COMPLETED | jonduke | 08/18/2019 11:21 PM | 08/18/2019 11:21 PM |
| 18 | Generating cohort characterization 1 : CMSDESynPUF1k (CMSDESynPUF1k) | COMPLETED | jonduke | 08/18/2019 11:21 PM | 08/18/2019 11:21 PM |
| 17 | Generating cohort 3 : CMSDESynPUF23m (CMSDESynPUF23m) | | jonduke | 08/18/2019 11:19 PM | 08/18/2019 11:19 PM |
| 16 | Generating cohort characterization 1 : CMSDESynPUF23m (CMSDESynP | | jonduke | 08/18/2019 11:19 PM | 08/18/2019 11:20 PM |
| 15 | Generating cohort 3 : CMSDESynPUF23m (CMSDESynPUF23m) | | jonduke | 08/18/2019 11:17 PM | 08/18/2019 11:17 PM |
| 14 | Generating cohort 3 : CMSDESynPUF23m (CMSDESynPUF23m) | COMPLETED | jonduke | 08/18/2019 11:16 PM | 08/18/2019 11:16 PM |
| 13 | Generating cohort 3 : CMSDESynPUF100k (CMSDESynPUF100k) | FAILED | jonduke | 08/18/2019 11:14 PM | 08/18/2019 11:14 PM |
| 12 | Generating cohort 3 : CMSDESynPUF1k (CMSDESynPUF1k) | COMPLETED | jonduke | 08/18/2019 11:13 PM | 08/18/2019 11:13 PM |
| 11 | warming CMSDESynPUF100k cache | COMPLETED | jonduke | 08/18/2019 8:04 PM | 08/18/2019 8:04 PM |
| 10 | warming CMSDESynPUF100k cache | COMPLETED | jonduke | 08/18/2019 8:03 PM | 08/18/2019 8:03 PM |
| 9 | Generating cohort 2 : CMSDESynPUF100k (CMSDESynPUF100k) | FAILED | jonduke | 08/18/2019 7:55 PM | 08/18/2019 7:55 PM |
| 8 | Generating cohort 2 : CMSDESynPUF1k (CMSDESynPUF1k) | COMPLETED | jonduke | 08/18/2019 7:54 PM | 08/18/2019 7:55 PM |
| 7 | Generating cohort 2 : CMSDESynPUF100k (CMSDESynPUF100k) | FAILED | jonduke | 08/18/2019 7:53 PM | 08/18/2019 7:53 PM |
| 6 | warming CMSDESynPUF23m cache | COMPLE | | | 8/2019 7:31 PM |
| 5 | warming CMSDESynPUF23m cache | COMPLE | | | 8/2019 7:26 PM |
| 4 | warming CMSDESynPUF23m cache | COMPLE | | | 7/2019 9:34 PM |
| 3 | warming CMSDESynPUF23m cache | COMPLETED | system | 08/17/2019 9:24 PM | 08/17/2019 9:26 PM |
| 2 | warming CMSDESynPUF23m cache | FAILED | system | 08/17/2019 9:21 PM | 08/17/2019 9:24 PM |
| 1 | warming CMSDESynPUF23m cache | COMPLETED | system | 08/17/2019 9:10 PM | 08/17/2019 9:12 PM |

As of 10/14/2019

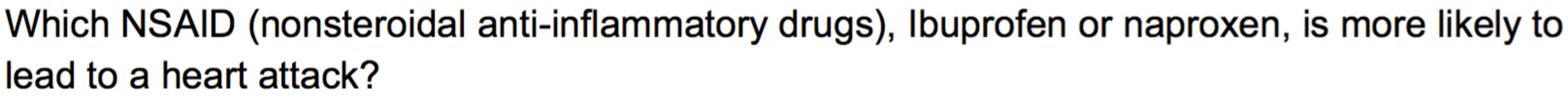
# Output

- In 7 weeks, the class generated
  - 2239 concept sets
  - 2343 cohorts
  - 825 characterizations
  - 905 incidence rates
  - 846 estimation studies
  - 386 study reports

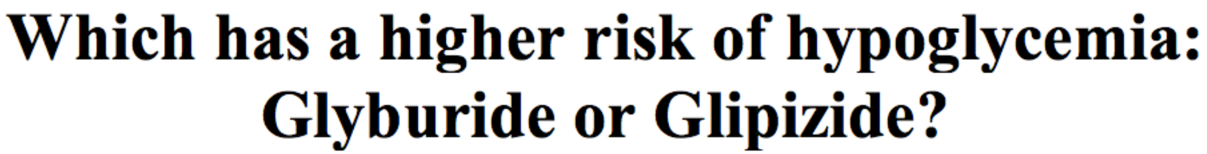
# Example Project Reports



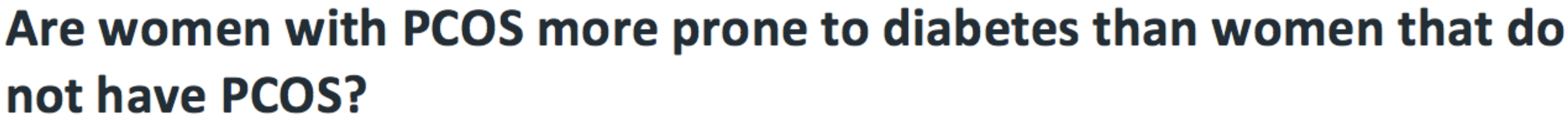
**An analysis of pulmonary embolism patients and the effect of heparin on their chance to develop thrombocytopenia.**



Which NSAID (nonsteroidal anti-inflammatory drugs), Ibuprofen or naproxen, is more likely to lead to a heart attack?



**Which has a higher risk of hypoglycemia: Glyburide or Glipizide?**



**Are women with PCOS more prone to diabetes than women that do not have PCOS?**

# What went well

- Students reported enjoying the chance to analyze data
  - Many students explored questions of personal interest
- Many students expressed interest in getting more engaged in OHDSI
- It was gratifying to see them help each other in solving problems and working through challenges

# Challenges

- We experienced a lot of challenges during the OHDSI block
- Although multi-factorial, I have categorized thematically
  - Vocabulary and concept set creation
  - Cohort definition
  - Running estimation studies
  - General infrastructure

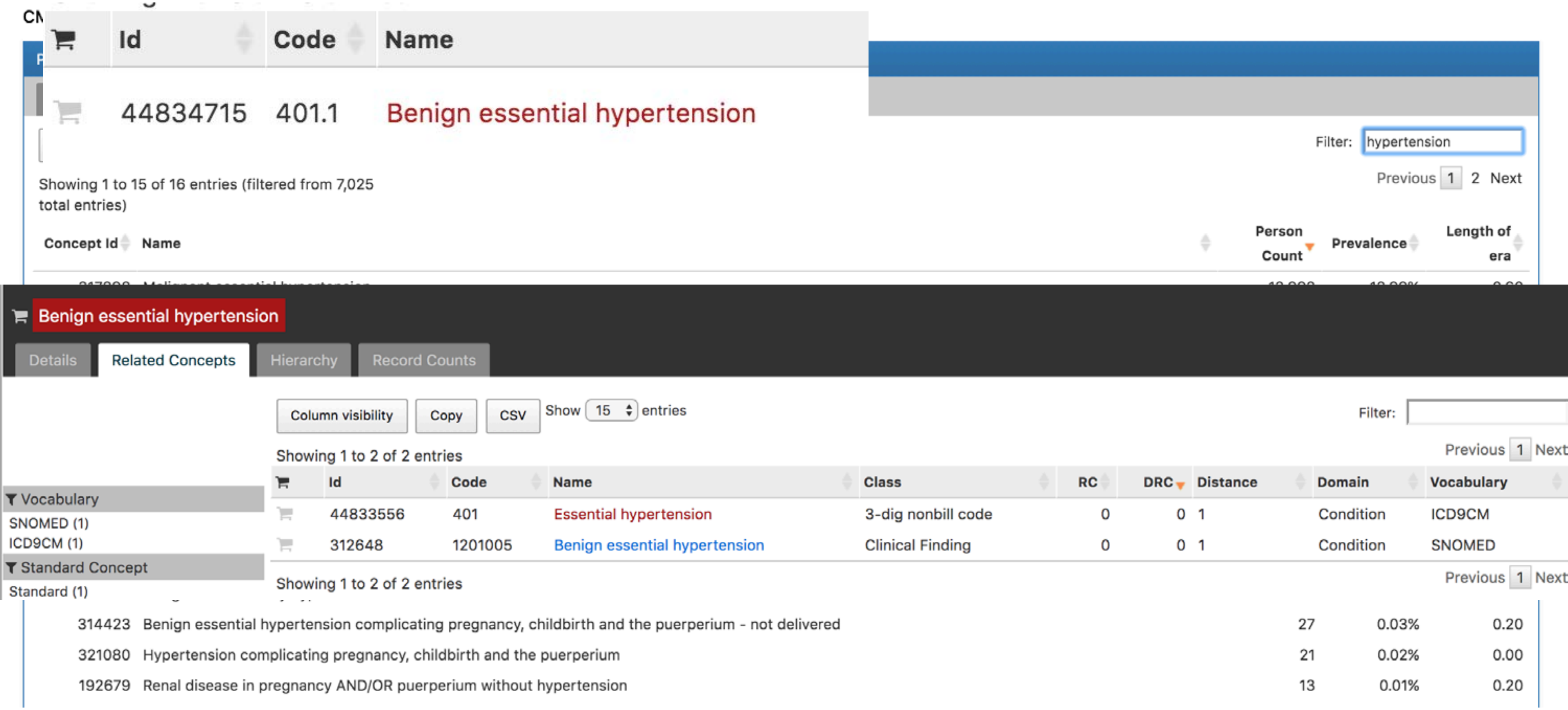
# Framing Potential Solutions

- For each challenge, I describe potential ideas
  - Note these do not distinguish things taking 5 minutes and things taking 5 months
- Solutions tagged as
  - Things I could have **taught better (T)**
  - Potential **software feature** enhancements **(S)**
  - OHDSI **Infrastructure (I)**

# Vocabulary and Concept Sets

- Finding standard concepts
  - Students were initially guided to find common ICD9/10 codes and use the OMOP vocabulary to find SNOMED codes
  - This was often not successful in the SynPUF dataset

# Example: Hypertension



| | Id | Code | Name |
|---|---|---|---|
| | 44834715 | 401.1 | Benign essential hypertension |

Filter: hypertension

Showing 1 to 15 of 16 entries (filtered from 7,025 total entries)

Previous 1 2 Next

| Concept Id | Name | Person Count | Prevalence | Length of era |
|---|---|---|---|---|
| 314423 | Benign essential hypertension complicating pregnancy, childbirth and the puerperium - not delivered | 27 | 0.03% | 0.20 |
| 321080 | Hypertension complicating pregnancy, childbirth and the puerperium | 21 | 0.02% | 0.00 |
| 192679 | Renal disease in pregnancy AND/OR puerperium without hypertension | 13 | 0.01% | 0.20 |

Benign essential hypertension

Show 15 entries

Filter:

Showing 1 to 2 of 2 entries

Previous 1 Next

Vocabulary
- SNOMED (1)
- ICD9CM (1)

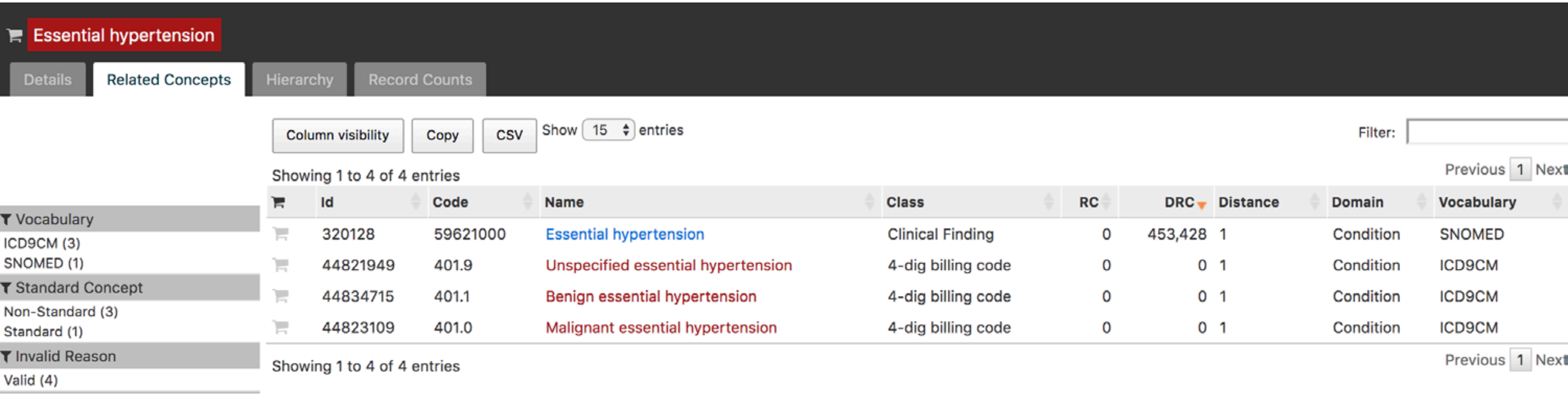
Standard Concept
- Standard (1)

| | Id | Code | Name | Class | RC | DRC | Distance | Domain | Vocabulary |
|---|---|---|---|---|---|---|---|---|---|
| | 44833556 | 401 | Essential hypertension | 3-dig nonbill code | 0 | 0 | 1 | Condition | ICD9CM |
| | 312648 | 1201005 | Benign essential hypertension | Clinical Finding | 0 | 0 | 1 | Condition | SNOMED |

Showing 1 to 2 of 2 entries

Previous 1 Next

# Had to search a level up to find



Vocabulary
ICD9CM (3)
SNOMED (1)

Standard Concept
Non-Standard (3)
Standard (1)

Invalid Reason
Valid (4)

Showing 1 to 4 of 4 entries

| | Id | Code | Name | Class | RC | DRC | Distance | Domain | Vocabulary |
|---|---|---|---|---|---|---|---|---|---|
| | 320128 | 59621000 | Essential hypertension | Clinical Finding | 0 | 453,428 | 1 | Condition | SNOMED |
| | 44821949 | 401.9 | Unspecified essential hypertension | 4-dig billing code | 0 | 0 | 1 | Condition | ICD9CM |
| | 44834715 | 401.1 | Benign essential hypertension | 4-dig billing code | 0 | 0 | 1 | Condition | ICD9CM |
| | 44823109 | 401.0 | Malignant essential hypertension | 4-dig billing code | 0 | 0 | 1 | Condition | ICD9CM |

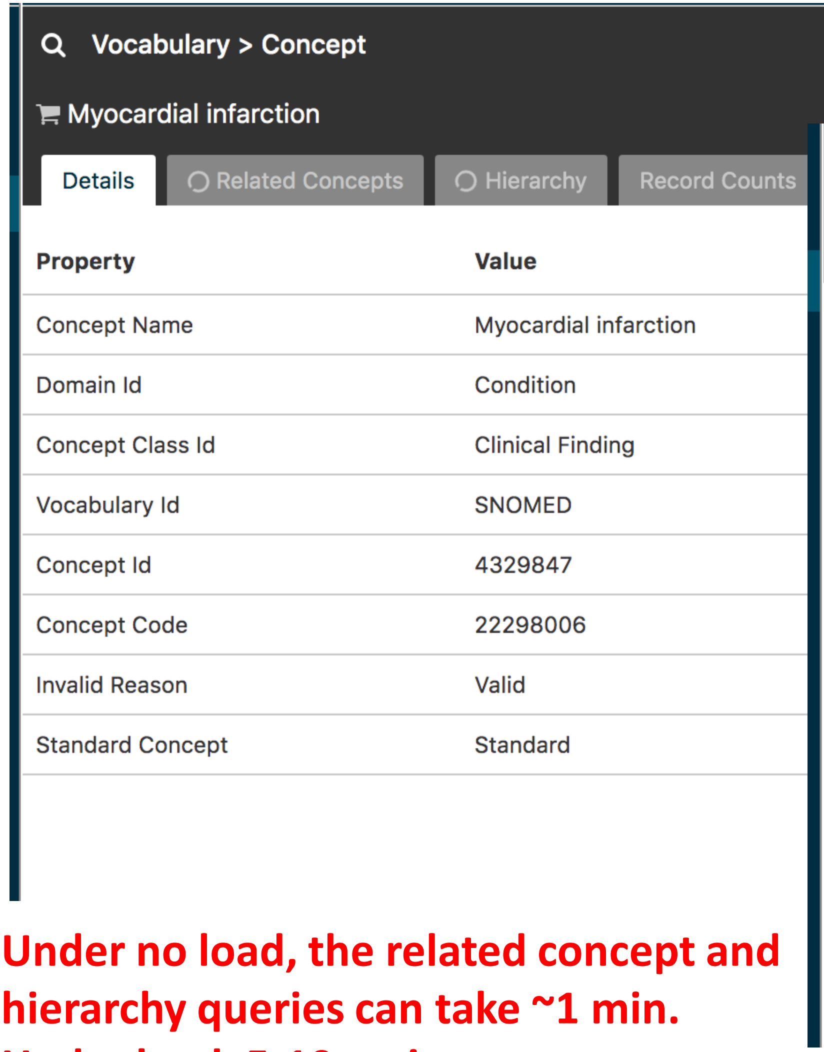
Showing 1 to 4 of 4 entries

But implications of DRC not sufficiently clear to students

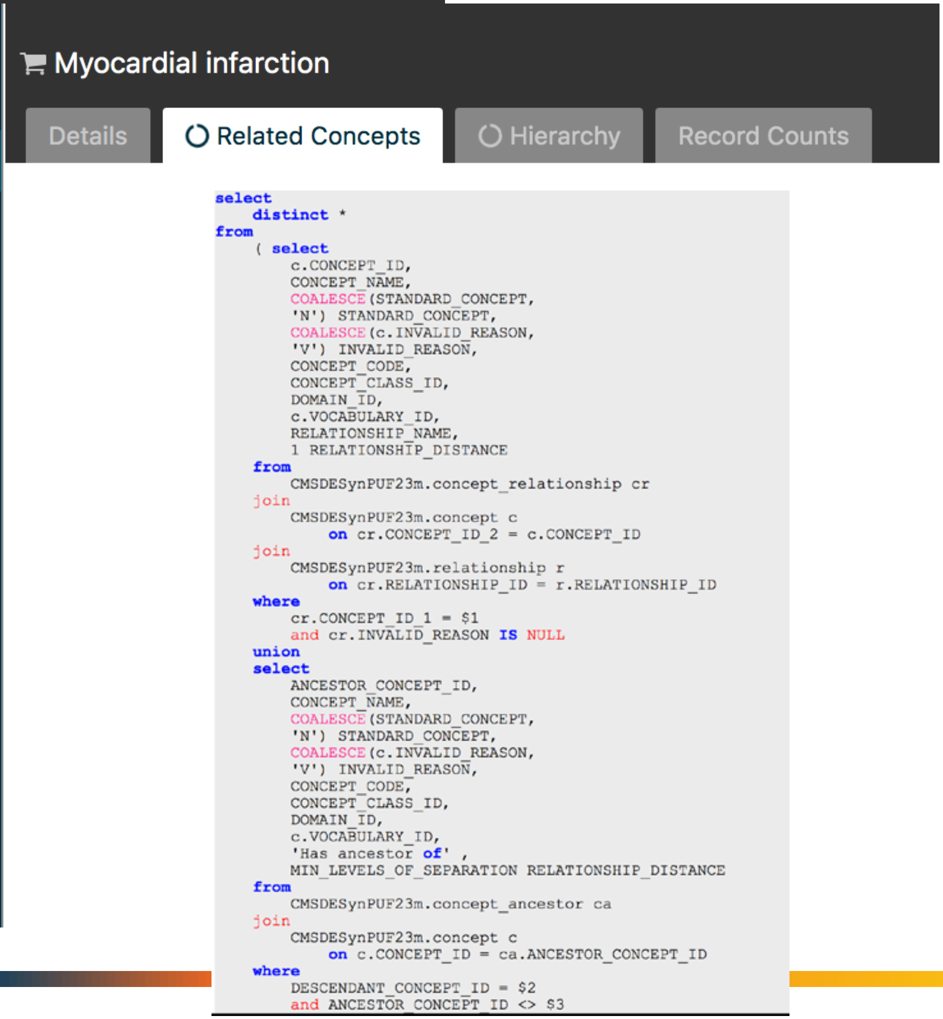
# DRC vs RC

- Sometimes students failed to select descendants and thus had 0 patients in cohort
- But use of descendants in concept sets carries its own problems in running Estimation studies (see section on Estimation Studies)

# The Most Expensive Query



| Property | Value |
|---|---|
| Concept Name | Myocardial infarction |
| Domain Id | Condition |
| Concept Class Id | Clinical Finding |
| Vocabulary Id | SNOMED |
| Concept Id | 4329847 |
| Concept Code | 22298006 |
| Invalid Reason | Valid |
| Standard Concept | Standard |

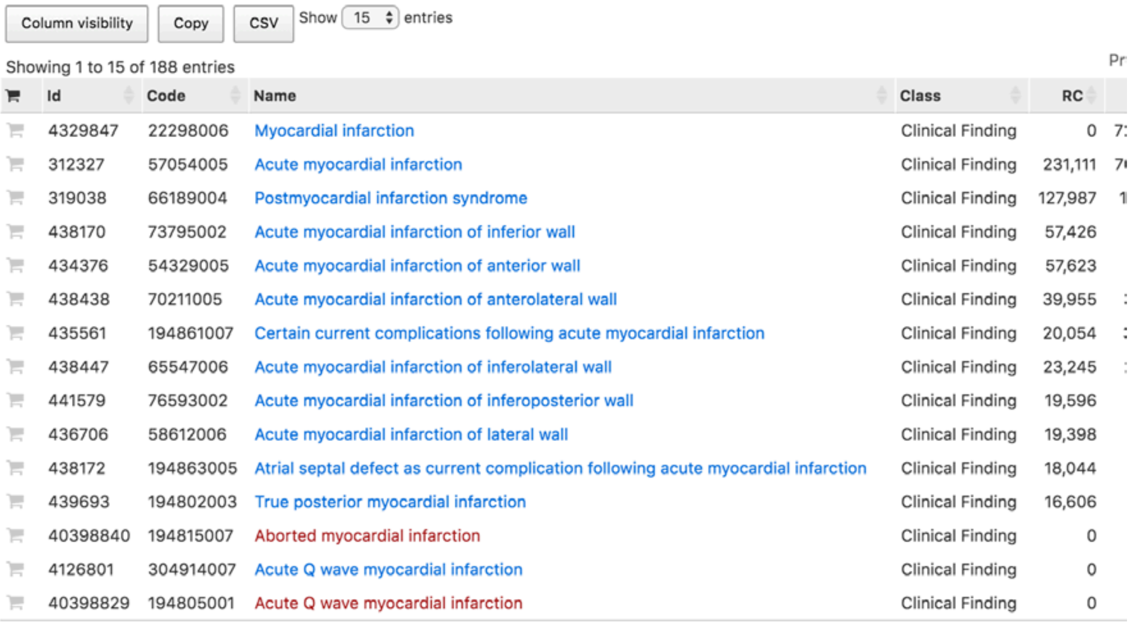


```sql
select
    distinct *
from
    ( select
        c.CONCEPT_ID,
        CONCEPT_NAME,
        COALESCE(STANDARD_CONCEPT,
        'N') STANDARD_CONCEPT,
        COALESCE(c.INVALID_REASON,
        'V') INVALID_REASON,
        CONCEPT_CODE,
        CONCEPT_CLASS_ID,
        DOMAIN_ID,
        c.VOCABULARY_ID,
        RELATIONSHIP_NAME,
        1 RELATIONSHIP_DISTANCE
    from
        CMSDESynPUF23m.concept_relationship cr
    join
        CMSDESynPUF23m.concept c
            on cr.CONCEPT_ID_2 = c.CONCEPT_ID
    join
        CMSDESynPUF23m.relationship r
            on cr.RELATIONSHIP_ID = r.RELATIONSHIP_ID
    where
        cr.CONCEPT_ID_1 = $1
        and cr.INVALID_REASON IS NULL
    union
    select
        ANCESTOR_CONCEPT_ID,
        CONCEPT_NAME,
        COALESCE(STANDARD_CONCEPT,
        'N') STANDARD_CONCEPT,
        COALESCE(c.INVALID_REASON,
        'V') INVALID_REASON,
        CONCEPT_CODE,
        CONCEPT_CLASS_ID,
        DOMAIN_ID,
        c.VOCABULARY_ID,
        'Has ancestor of' ,
        MIN_LEVELS_OF_SEPARATION RELATIONSHIP_DISTANCE
    from
        CMSDESynPUF23m.concept_ancestor ca
    join
        CMSDESynPUF23m.concept c
            on c.CONCEPT_ID = ca.ANCESTOR_CONCEPT_ID
    where
        DESCENDANT_CONCEPT_ID = $2
        and ANCESTOR_CONCEPT_ID <> $3
```

**Under no load, the related concept and hierarchy queries can take ~1 min. Under load, 5-10+ mins**

# The Most Expensive Query

- These are not rare queries, as they are run automatically every time any concept is clicked



Showing 1 to 15 of 188 entries

Pr

| | Id | Code | Name | Class | RC | |
|---|---|---|---|---|---|---|
| | 4329847 | 22298006 | Myocardial infarction | Clinical Finding | 0 | 7: |
| | 312327 | 57054005 | Acute myocardial infarction | Clinical Finding | 231,111 | 7 |
| | 319038 | 66189004 | Postmyocardial infarction syndrome | Clinical Finding | 127,987 | 1 |
| | 438170 | 73795002 | Acute myocardial infarction of inferior wall | Clinical Finding | 57,426 | |
| | 434376 | 54329005 | Acute myocardial infarction of anterior wall | Clinical Finding | 57,623 | |
| | 438438 | 70211005 | Acute myocardial infarction of anterolateral wall | Clinical Finding | 39,955 | : |
| | 435561 | 194861007 | Certain current complications following acute myocardial infarction | Clinical Finding | 20,054 | : |
| | 438447 | 65547006 | Acute myocardial infarction of inferolateral wall | Clinical Finding | 23,245 | : |
| | 441579 | 76593002 | Acute myocardial infarction of inferoposterior wall | Clinical Finding | 19,596 | |
| | 436706 | 58612006 | Acute myocardial infarction of lateral wall | Clinical Finding | 19,398 | |
| | 438172 | 194863005 | Atrial septal defect as current complication following acute myocardial infarction | Clinical Finding | 18,044 | |
| | 439693 | 194802003 | True posterior myocardial infarction | Clinical Finding | 16,606 | |
| | 40398840 | 194815007 | Aborted myocardial infarction | Clinical Finding | 0 | |
| | 4126801 | 304914007 | Acute Q wave myocardial infarction | Clinical Finding | 0 | |
| | 40398829 | 194805001 | Acute Q wave myocardial infarction | Clinical Finding | 0 | |

# Concept Set Creation

- Ended up recommending that most people utilize **Atlas Data Sources** (ie ACHILLES) to find the concepts actually present in the dataset instead of using vocabulary-based lookup
  - Some exceptions for broad outcomes with many descendants (eg Cancer)
- Use of RxNorm ingredients vs Clinical Drugs was also not well-grokked by many student so did similar thing for drug era concepts

# Potential Solutions

- More didactic time dedicated to DRC vs RC, RxNorm components (T)
- Change Atlas trigger for WebAPI call for related concepts and hierarchy to clicking on tabs (S)
- Reviewing DB query optimization strategies for vocabulary based queries (I)

# Cohort Generation

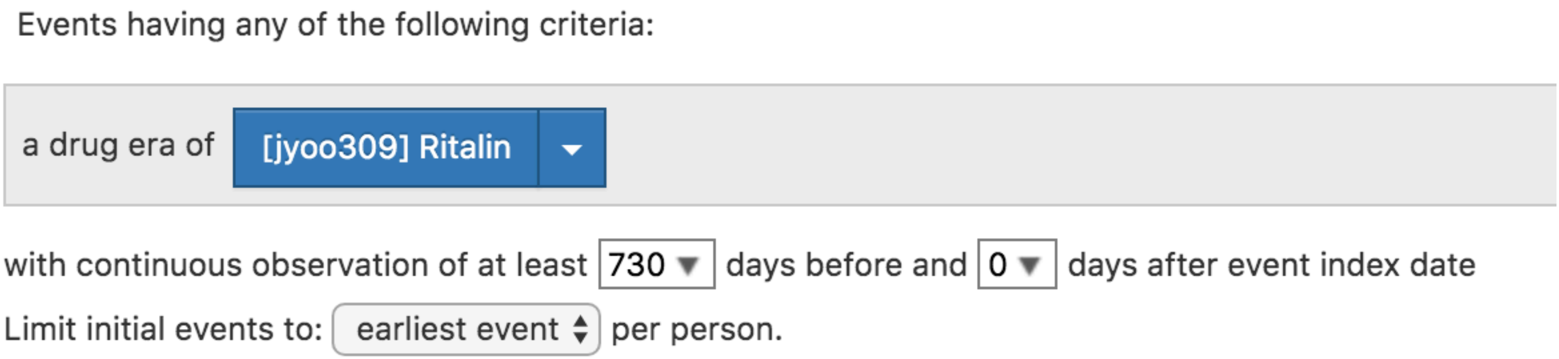
- Cohorts had two flavors of problems
  - Cohorts that intrinsically fail to produce patients
  - Cohort that produce patients but are not well aligned with conducting an estimation study

# Failing to produce patients

- Problems with concept sets as above
- Required continuous observation period excessively long for SynPUF (2 yrs total data)
- Despite extensive discussion on claims databases and SynPUF, still a lot of pediatric, OB, etc cohorts trying to be generated

# Failing to produce patients

- Problems with concept sets as above
- Required continuous observation period excessively long for SynPUF (2 yrs total data)



Events having any of the following criteria:

a drug era of [jyoo309] Ritalin

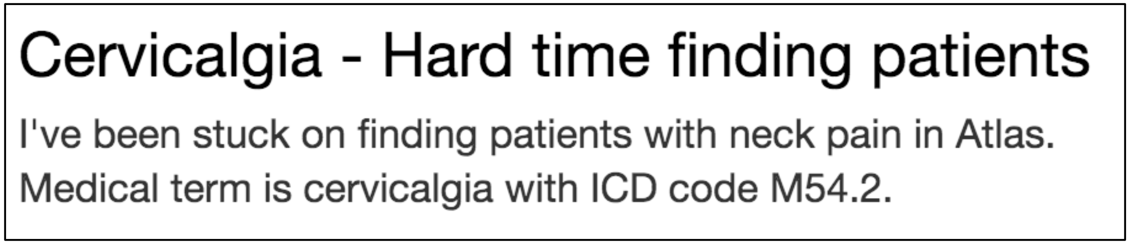
with continuous observation of at least 730 days before and 0 days after event index date

Limit initial events to: earliest event per person.

# Failing to produce patients

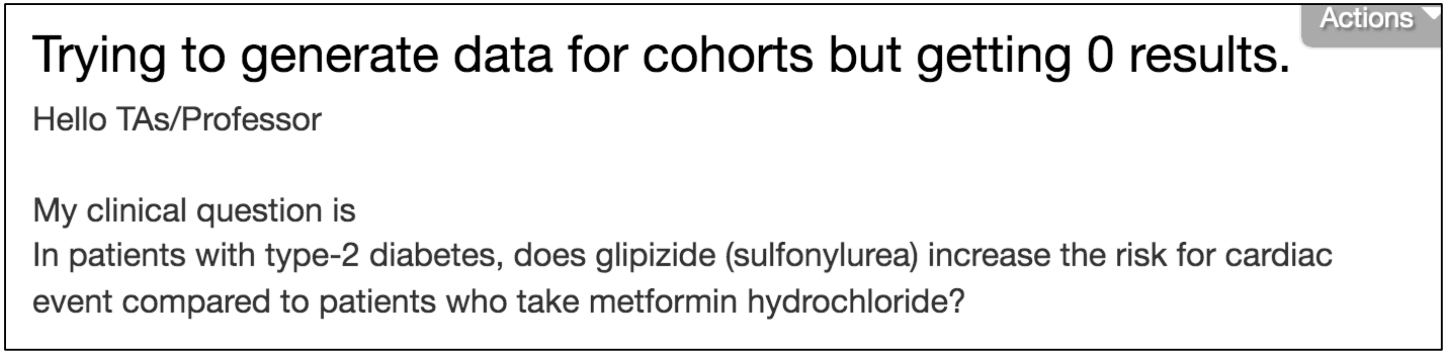
- Problems with concept sets as above
- Required continuous observation period excessively long for SynPUF (2 yrs total data)
- Despite extensive discussion on claims databases and SynPUF, still a lot of pediatric, OB, etc cohorts trying to be generated

# Zero Patient Blues



## Cervicalgia - Hard time finding patients

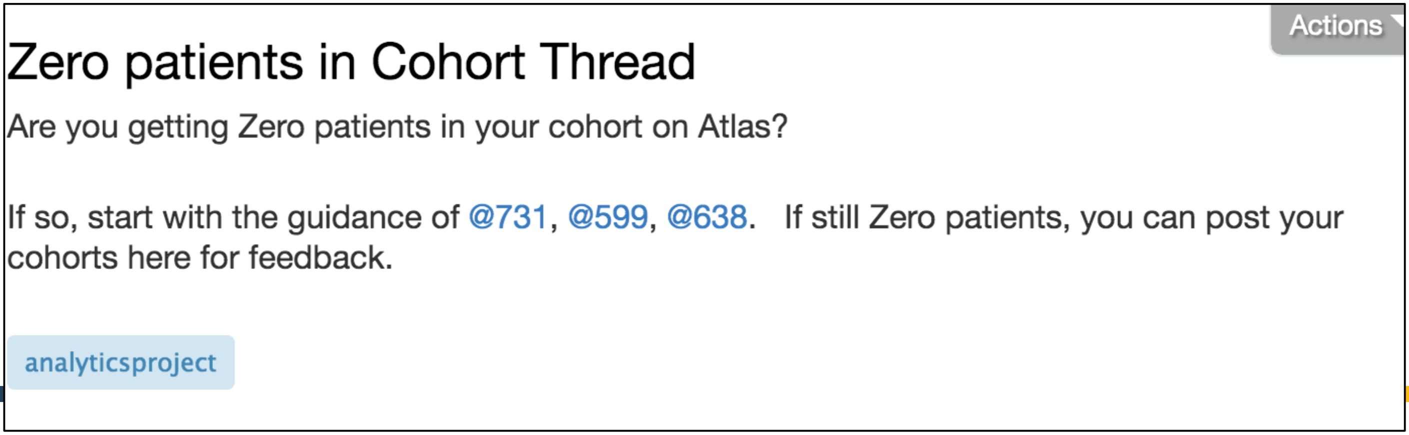
I've been stuck on finding patients with neck pain in Atlas. Medical term is cervicalgia with ICD code M54.2.



## Trying to generate data for cohorts but getting 0 results.

Hello TAs/Professor

My clinical question is
In patients with type-2 diabetes, does glipizide (sulfonylurea) increase the risk for cardiac event compared to patients who take metformin hydrochloride?



## Zero patients in Cohort Thread

Are you getting Zero patients in your cohort on Atlas?

If so, start with the guidance of @731, @599, @638. If still Zero patients, you can post your cohorts here for feedback.

analyticsproject

# Cohorts that Fail in Estimation Studies

- With tips on concept finding and temporal settings, most students were able to generate populated cohorts and successfully run characterization and incidence rates in Atlas
- But many students who were able to produce T, C, and O cohorts and reasonable incidence rates were still unable to successfully run Estimation Studies

```
  variate balanc

  od 1 returns error: "unab         when using ar
   list(analysisId = 1, targe      34, comparatorId
  comeId = 836, includedCovari     nceptIds = 1, exc
  iateConceptIds = 1, outcomeOf   erest = TRUE, coho
  aFolder = "CmData_l1_t834_c     , studyPopFile =
  1_s1_t834_c835_o836.rds", s    edPsFile = "Ps_l1_
  c835.rds", psFile = "Ps_l1_s  1_t834_c835_o836.r
ds", strataFile = "StratPop_l1_s1_p1_t8  c835_s1_o836.rds",
 prefilteredCovariatesFolder = "", outc  ModelFile = "Analy
sis_1/om_t834_c835_o836.rds"),~/MyProje  /EstimationResult
s/FertilityStudy/cmOutput,~/MyProjects/  timationResults/Fer
tilityStudy/balance
```

# Estimation Study Errors

- Many studies failed in the **compute covariate balance** phase
- After investigation (thanks Jamie Weaver!), these errors were typically due to:
  - Insufficient prior observation period, often requiring 365 days of pre-index to compute
  - T and C cohorts too divergent (comparator cohort not an ‘active comparator’, just too different)
  - T / C cohort too small for any matched patients to emerge from PS-score matching process
  - Covariate exclusion concept sets included descendants, whereas CohortMethod prefers parent concepts only accompanied by ”include descendants” in study design

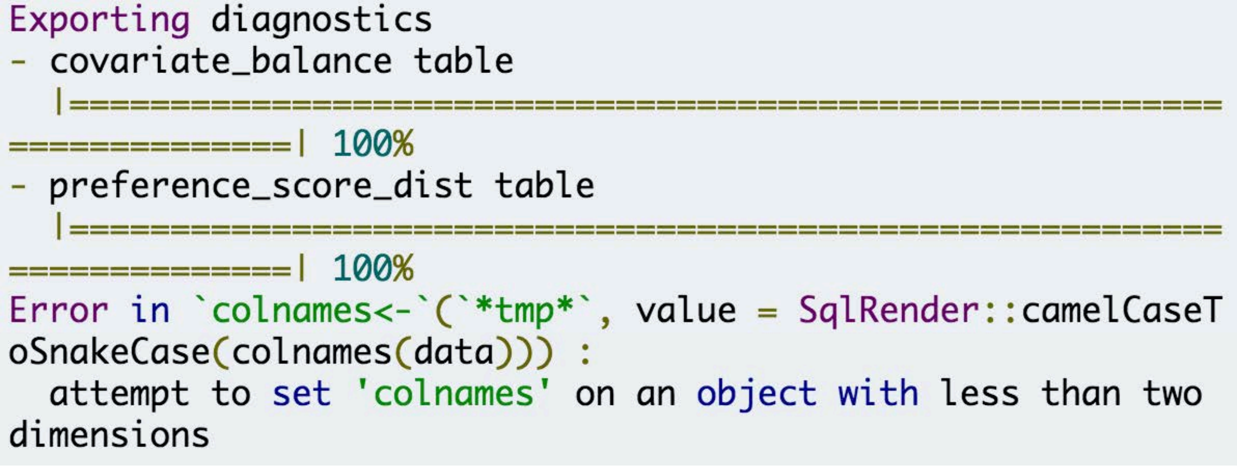
```
variate balance
od 1 returns error: "unab          when using ar
  list(analysisId = 1, targe      34, comparatorId
comeId = 836, includedCovari      nceptIds = 1, exc
iateConceptIds = 1, outcomeO      erest = TRUE, coho
aFolder = "CmData_l1_t834_c      , studyPopFile =
l1_s1_t834_c835_o836.rds",        edPsFile = "Ps_l1_
c835.rds", psFile = "Ps_l1_s      1_t834_c835_o836.r
ds", strataFile = "StratPop_l1_s1_p1_t8   c835_s1_o836.rds",
 prefilteredCovariatesFolder = "", outc   ModelFile = "Analy
sis_1/om_t834_c835_o836.rds"),~/MyProje   /EstimationResult
s/FertilityStudy/cmOutput,~/MyProjects/   timationResults/Fer
tilityStudy/balance
```

# Estimation Study Errors

- Some studies achieved patient matching but ended up with zero outcomes
  - This was often due to outcome cohort observation period requirements being too long for SynPUF
  - Or just small numbers of patients with the chosen outcome so matching ended up at zero
- MethodEvaluation will error if zero outcomes so cannot use Shiny app to view output on cohorts, covariate balance, etc

# Estimation Study Errors

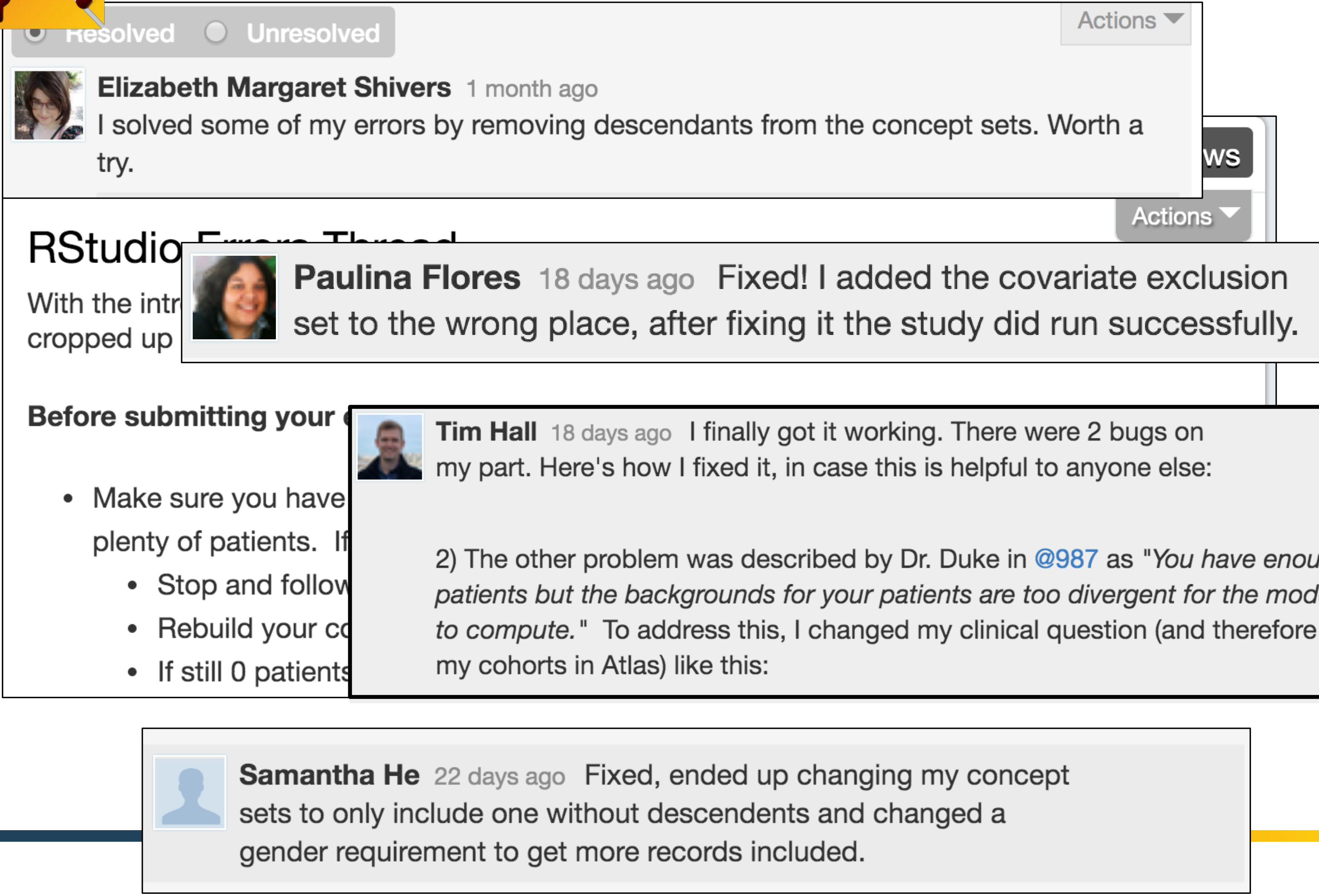
- Some studies failed in the Export phase with the mysterious camelCaseToSnakeCase error



```
Exporting diagnostics
- covariate_balance table
  |=========================================================
==============|  100%
- preference_score_dist table
  |=========================================================
==============|  100%
Error in `colnames<-`(`*tmp*`, value = SqlRender::camelCaseT
oSnakeCase(colnames(data))) : 
  attempt to set 'colnames' on an object with less than two
dimensions
```

- This is due to T and C cohorts being *so similar* that all patients are assigned a propensity of 0.5 for every covariate

# Active Discussion on these Topics



**Elizabeth Margaret Shivers** 1 month ago
I solved some of my errors by removing descendants from the concept sets. Worth a try.

RStudio Errors Thread

With the intr
cropped up

**Before submitting your**

- Make sure you have plenty of patients. If
  - Stop and follow
  - Rebuild your co
  - If still 0 patients

**Paulina Flores** 18 days ago Fixed! I added the covariate exclusion set to the wrong place, after fixing it the study did run successfully.

**Tim Hall** 18 days ago I finally got it working. There were 2 bugs on my part. Here's how I fixed it, in case this is helpful to anyone else:

2) The other problem was described by Dr. Duke in @987 as "*You have enou patients but the backgrounds for your patients are too divergent for the mod to compute.*" To address this, I changed my clinical question (and therefore my cohorts in Atlas) like this:

**Samantha He** 22 days ago Fixed, ended up changing my concept sets to only include one without descendents and changed a gender requirement to get more records included.

# Active Comparators Can Be Hard to Come By

- Picking a good active comparator takes some clinical informatics knowledge, so setting 400 CS students loose on their own questions with just one Dr. Duke was, in retrospect, unwise
- That said, it is hard to find a clinically accurate active comparator for many questions that real people ask, eg
  - Do women who get mammograms have a lower risk of breast cancer than women who don't?
  - Do women with PCOS have a higher risk for diabetes than women without PCOS?
  - Does long-term antibiotic use increase risk for myocardial infarction?

Does Zantac cause alopecia?

Compared to what?

People who don’t take Zantac.

Not a good comparator.

Men who don’t take Zantac.

Not a good comparator.

Men with GERD who don’t take Zantac.

Not a good comparator.

Men with GERD who were given Prilosec?

Great study!

Umm, that wasn’t my question...

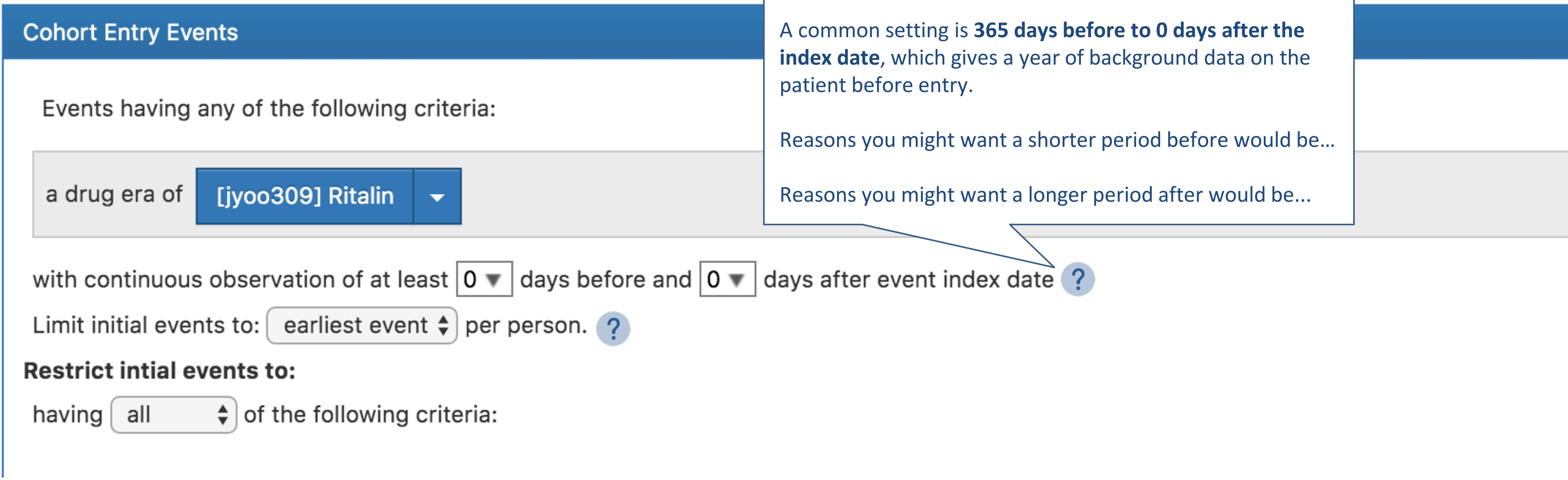
# Waxing Philosophically for a Moment

- CohortMethod is designed to perform a particular task– to compare a cohort X with active comparator cohort Y for viable outcome O in a database with sufficient patients to answer this
- It is a valid question of whether
  - I need to teach my students how to better design their questions to match CohortMethod expectations
  - OHDSI needs additional packages and/or guidance in our tools to allow people to answer basic (non study-grade) questions without running aground on errors

# Waxing Philosophically for a Moment

- Likely a hybrid approach of expanded didactics, more guidance around errors, and additions to Atlas would bridge the gap
  - Atlas is extremely powerful and can produce almost everything you need for a good first look at a question (characterization, incidence)
  - Temporality is a killer, though, particularly for smaller databases, so maybe including **decision support around cohort design** that could help users understand implication of time restrictions with their data

# Example Support in Atlas



Cohort Entry Events

Events having any of the following criteria:

a drug era of [jyoo309] Ritalin ▾

with continuous observation of at least 0 ▾ days before and 0 ▾ days after event index date ?

Limit initial events to: earliest event ⬍ per person. ?

**Restrict intial events to:**

having all ⬍ of the following criteria:

Continuous observation period sets the duration the patient must be present in the dataset in order for the index event to match.

A common setting is **365 days before to 0 days after the index date**, which gives a year of background data on the patient before entry.

Reasons you might want a shorter period before would be…

Reasons you might want a longer period after would be…

# Some Ideas

- More teaching on Active Comparators (T)
- Fixes to Atlas / PLE to clean up complications around descendants, exclusion set location (S)
- Cohort templates on OHDSI.org for how to answer certain kinds of common questions (T/S)
- Estimation templates on OHDSI.org with “liberal” study parameters (T/S)
- Kaplan-Meier curve in Atlas (S)
- More informative errors in study package (S)

# Infrastructure

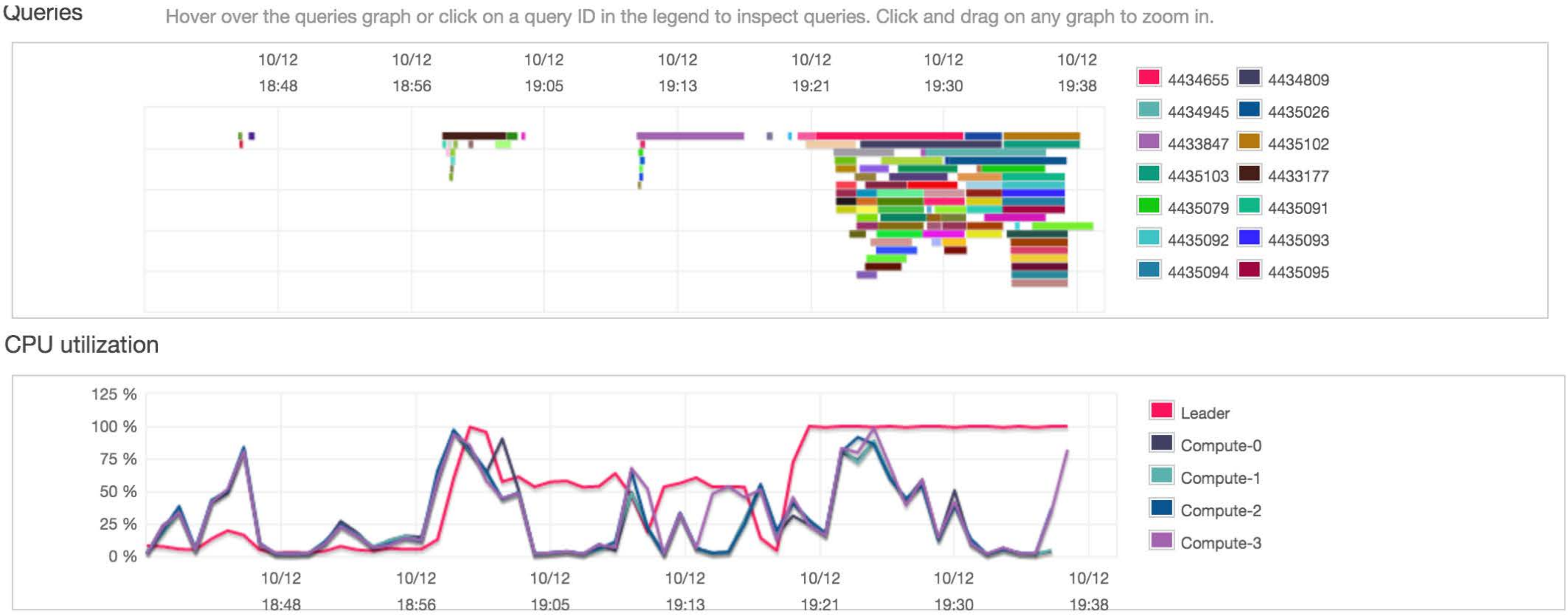
designed by freepik

# RStudio

- Robust, stable, handled student load well
- With so many studies, did have problems with tmp folder filling up and crashing things
- But overall super stable

# SynPUF OMOP CDM on Redshift

- Most queries (previous vocabulary exceptions noted) ran very fast under low user load
- But increased load really slowed things down

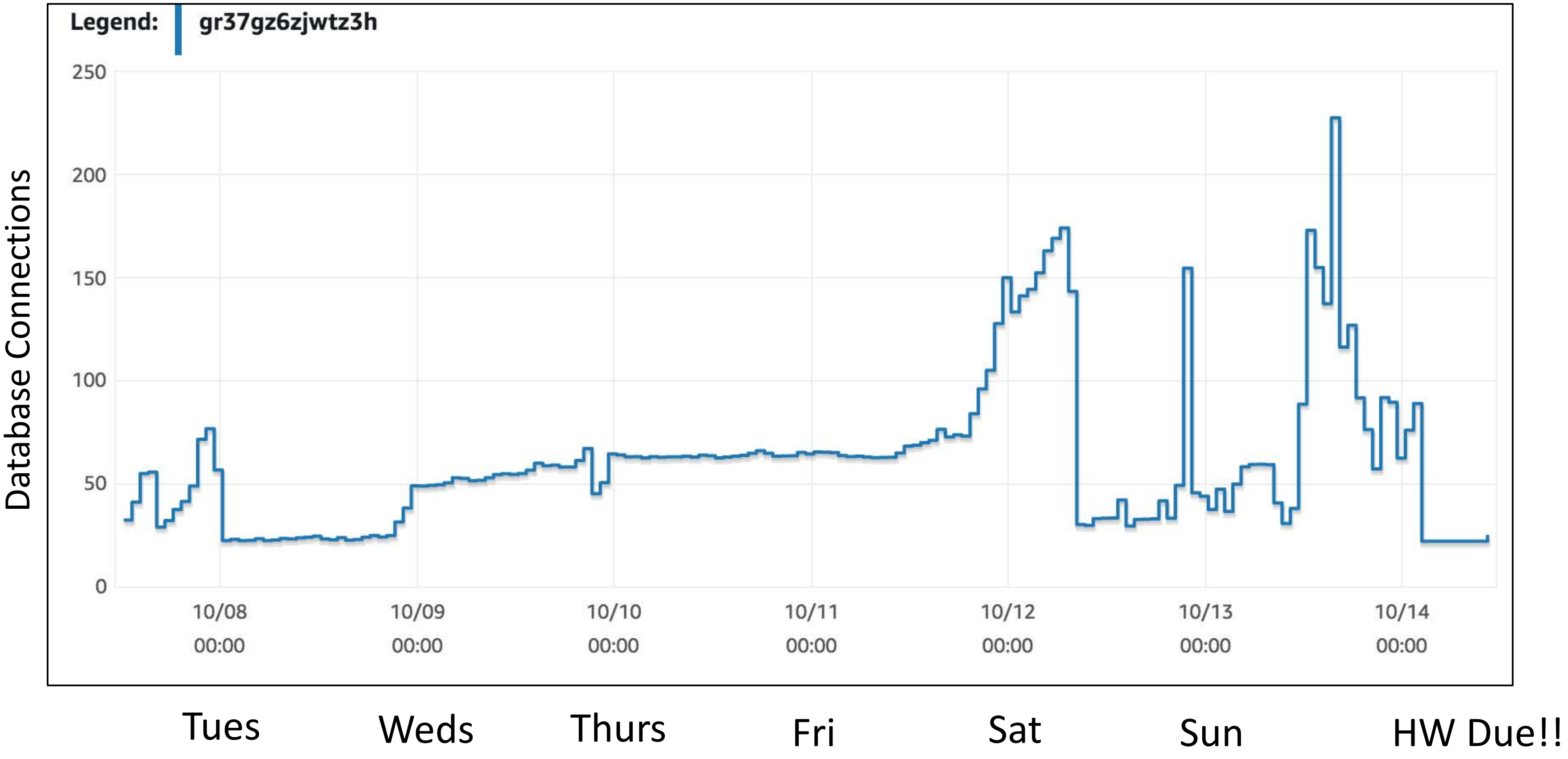


Queries

Hover over the queries graph or click on a query ID in the legend to inspect queries. Click and drag on any graph to zoom in.

CPU utilization

# What was the DB load?



# Atlas / WebAPI

- The OHDSI ecosystem is of course many systems running together
- But as the ‘tip of the spear’, Atlas bore the brunt of the stability issues and ire from students
- Despite 2-4 nodes on Elastic Beanstalk, it required frequent rebooting to address issues of very slow or failing jobs under load

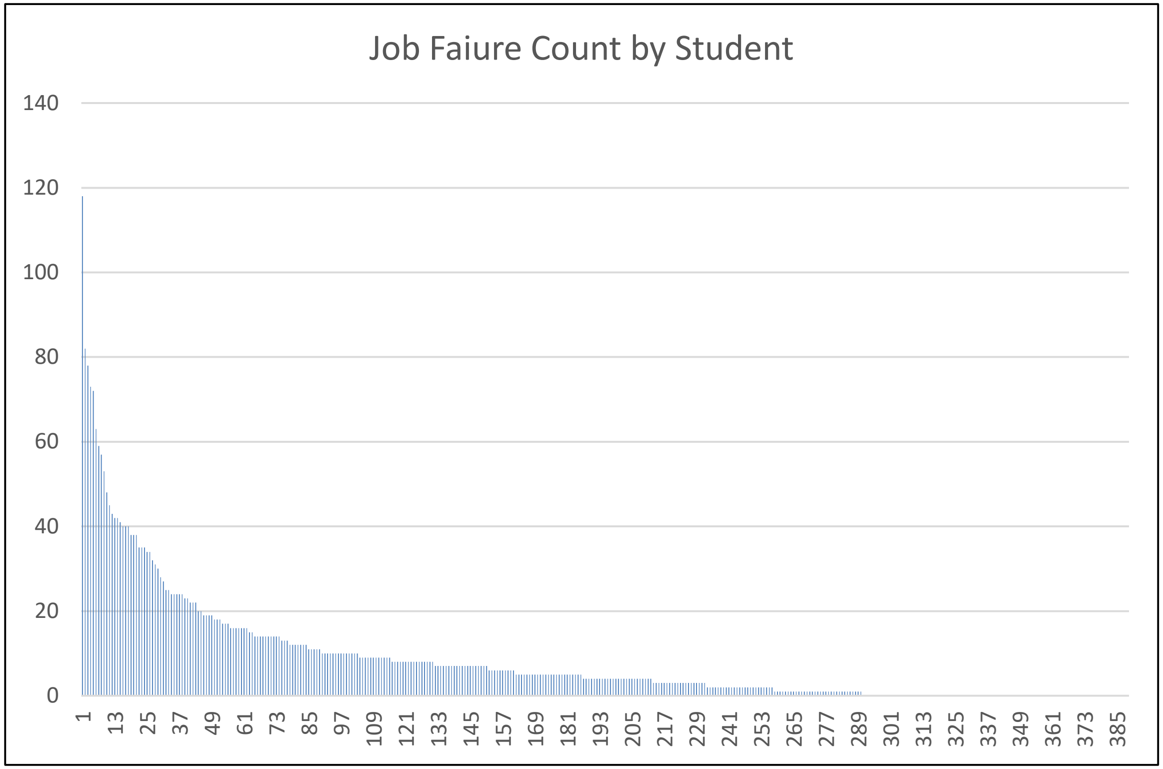
# Atlas Job Performance

| Type of Job | Proportion of Total |
|---|---|
| Cohort Generation | 81.07% |
| Incidence Rate | 12.04% |
| Characterization | 5.30% |
| Other (eg cache) | 1.59% |

| Type of Job | COMPLETED | FAILED | STARTING | STOPPED | STOPPING |
|---|---|---|---|---|---|
| Cohort | 93.62% | 1.84% | 4.02% | 0.49% | 0.02% |
| IR | 86.31% | 3.50% | 4.62% | 5.49% | 0.00% |
| Characterization | 78.51% | 18.48% | 0.00% | 3.01% | 0.00% |
| Other (eg cache) | 84.30% | 11.13% | 3.96% | 0.00% | 0.00% |
| **Overall** | **91.79%** | **3.07%** | **3.88%** | **1.22%** | **0.02%** |

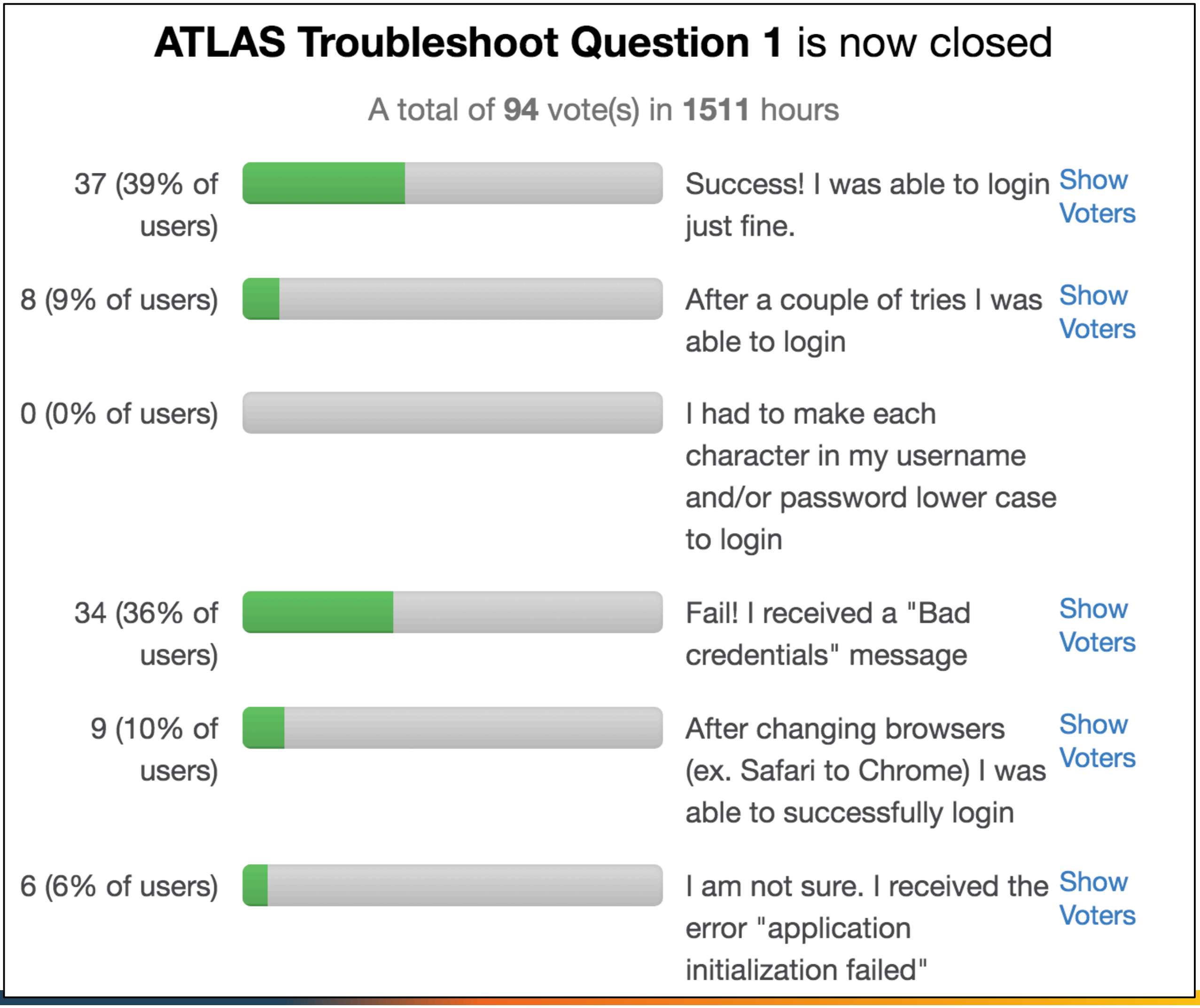
# Atlas Job Performance

- 74% of students experienced at least one failed job (range 1 to 118 failures per student)



# Atlas Authentication

Some students had trouble logging into Atlas initially



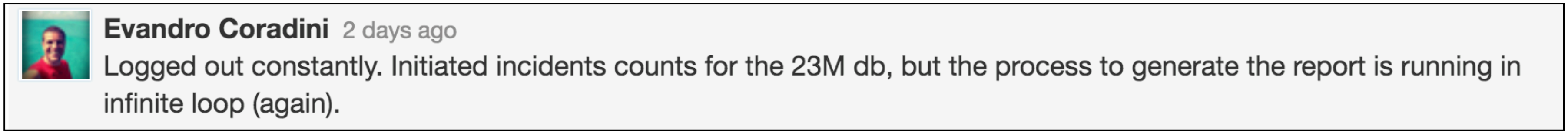
**ATLAS Troubleshoot Question 1** is now closed

A total of **94** vote(s) in **1511** hours

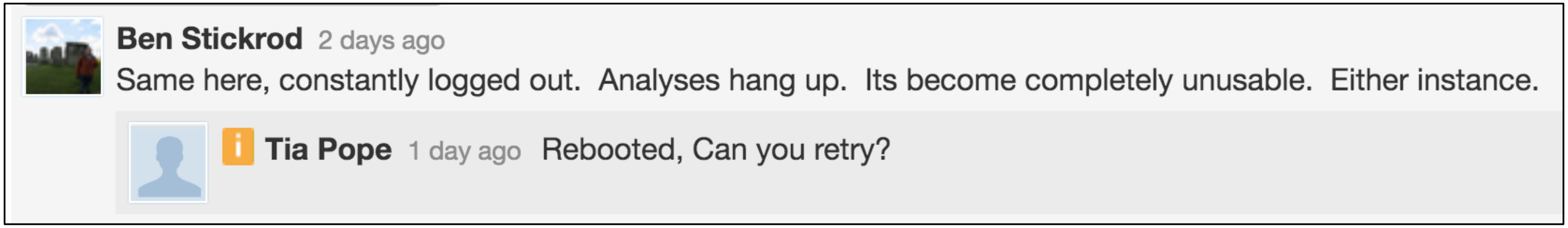
| | | |
|---|---|---|
| 37 (39% of users) | Success! I was able to login just fine. | Show Voters |
| 8 (9% of users) | After a couple of tries I was able to login | Show Voters |
| 0 (0% of users) | I had to make each character in my username and/or password lower case to login | |
| 34 (36% of users) | Fail! I received a "Bad credentials" message | Show Voters |
| 9 (10% of users) | After changing browsers (ex. Safari to Chrome) I was able to successfully login | Show Voters |
| 6 (6% of users) | I am not sure. I received the error "application initialization failed" | Show Voters |

# Atlas Authentication

- Subsequently issues possibly related to sticky sessions or server reboots led many students to experience **frequent logouts** by the system



**Evandro Coradini** 2 days ago
Logged out constantly. Initiated incidents counts for the 23M db, but the process to generate the report is running in infinite loop (again).

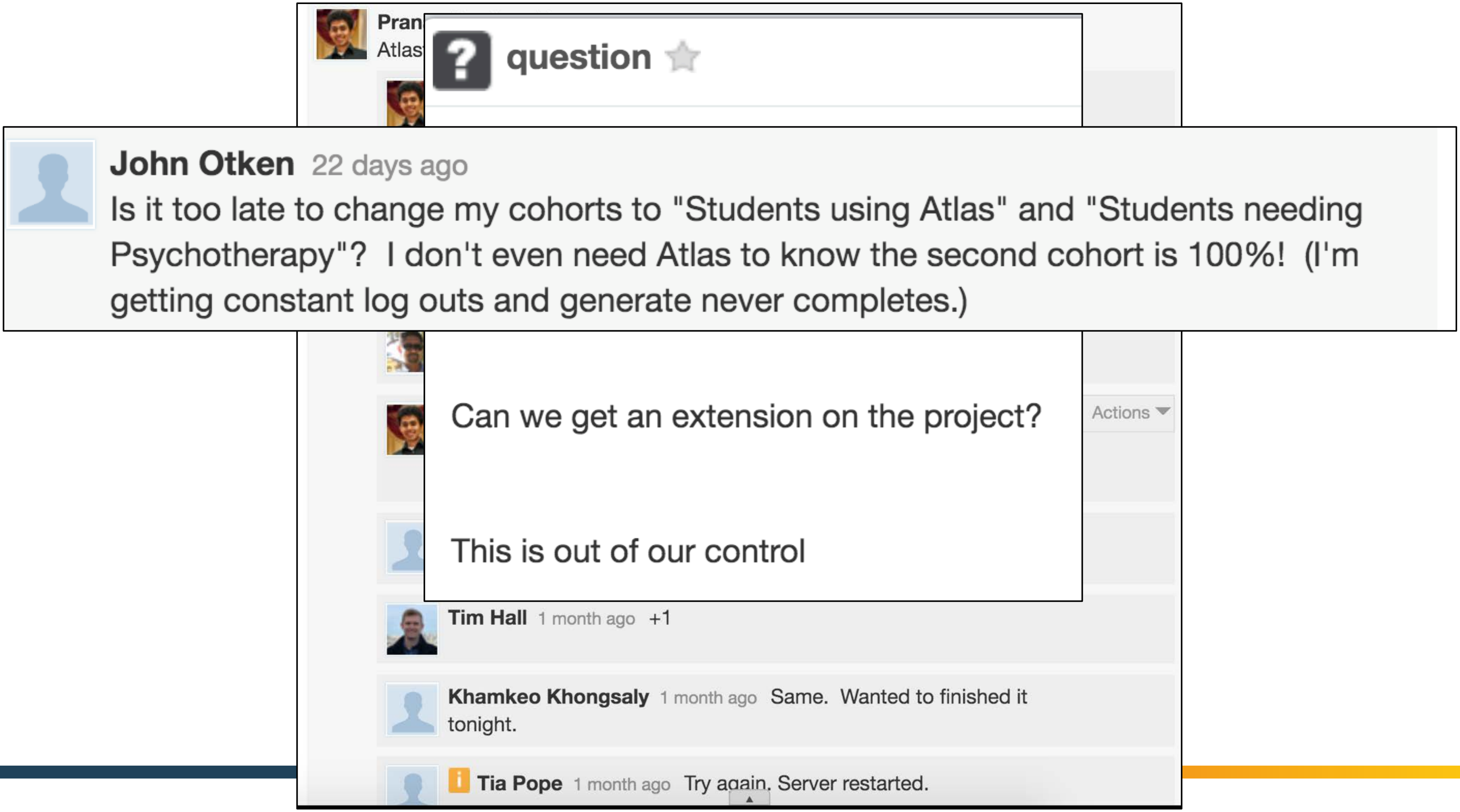


**Ben Stickrod** 2 days ago
Same here, constantly logged out. Analyses hang up. Its become completely unusable. Either instance.

**Tia Pope** 1 day ago Rebooted, Can you retry?

# Atlas / WebAPI

- Atlas (and I) took some heat from the students



Pran

Atlas

question

John Otken 22 days ago

Is it too late to change my cohorts to "Students using Atlas" and "Students needing Psychotherapy"? I don't even need Atlas to know the second cohort is 100%! (I'm getting constant log outs and generate never completes.)

Can we get an extension on the project?

This is out of our control

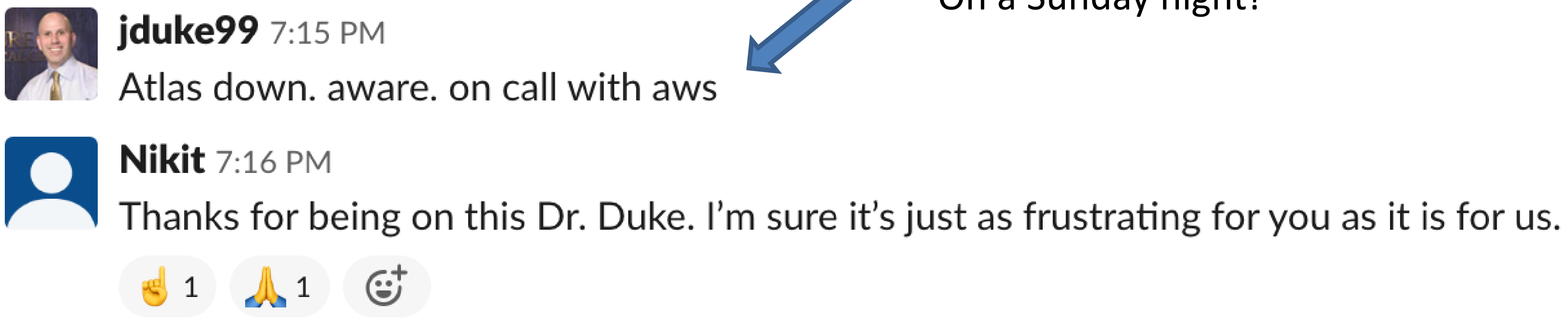
Tim Hall 1 month ago +1

Khamkeo Khongsaly 1 month ago Same. Wanted to finished it tonight.

Tia Pope 1 month ago Try again. Server restarted.

# But the OHDSI community is always there to lend a hand…

**James Wiggins!** 🙏 1
On a Sunday night!



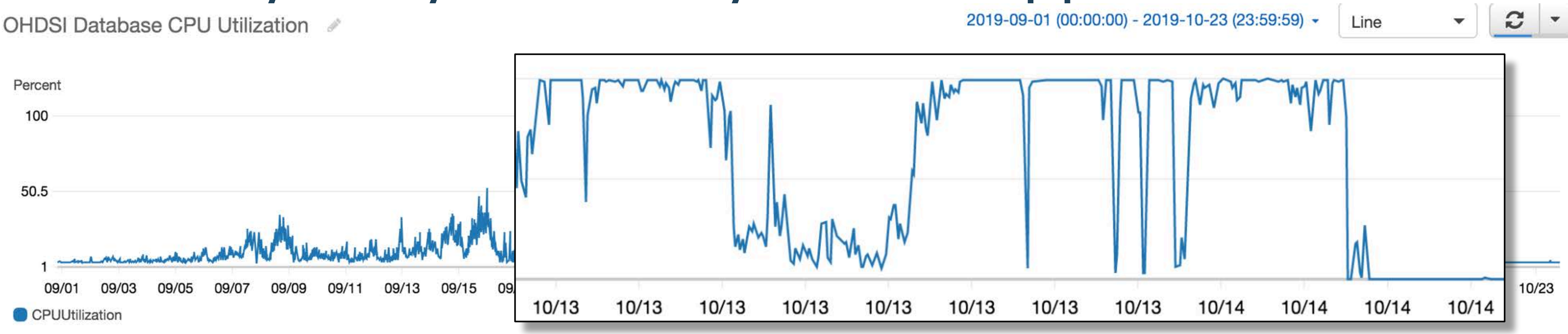
**jduke99** 7:15 PM
Atlas down. aware. on call with aws

**Nikit** 7:16 PM
Thanks for being on this Dr. Duke. I’m sure it’s just as frustrating for you as it is for us.

☝ 1 🙏 1

# Possible Explanations

- My sense is that the Atlas issues were not due primarily to OMOP CDM database issues
- The number of users and number of jobs may have exacerbated existing small memory leaks
- But some cumulative effect was seen on the OHDSI PG database over the 6 weeks, which is likely a key factor beyond the application



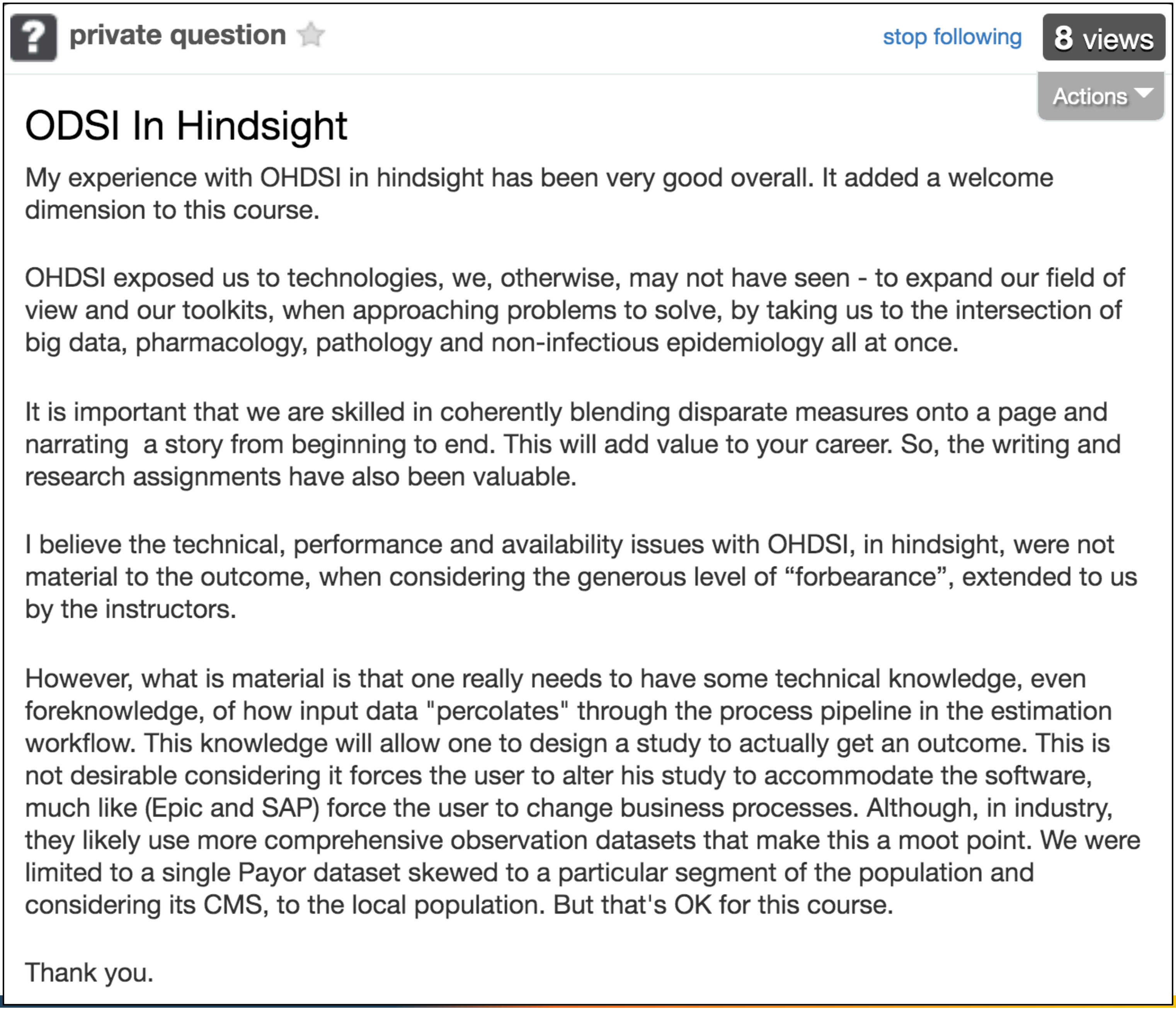
# Potential Solutions

- Don’t run classes with 400 online students having midnight deadlines (T)
- As OHDSI looks towards Atlas 3.0, good opportunity to leverage the ever-growing technical expertise for enhancements to (I)
  - job/pipeline management
  - memory management
  - load testing
  - Other great things I have no idea about

# So…

or

Received several notes from students re OHDSI. Here’s my favorite.



## ODSI In Hindsight

My experience with OHDSI in hindsight has been very good overall. It added a welcome dimension to this course.

OHDSI exposed us to technologies, we, otherwise, may not have seen - to expand our field of view and our toolkits, when approaching problems to solve, by taking us to the intersection of big data, pharmacology, pathology and non-infectious epidemiology all at once.

It is important that we are skilled in coherently blending disparate measures onto a page and narrating a story from beginning to end. This will add value to your career. So, the writing and research assignments have also been valuable.

I believe the technical, performance and availability issues with OHDSI, in hindsight, were not material to the outcome, when considering the generous level of “forbearance”, extended to us by the instructors.

However, what is material is that one really needs to have some technical knowledge, even foreknowledge, of how input data "percolates" through the process pipeline in the estimation workflow. This knowledge will allow one to design a study to actually get an outcome. This is not desirable considering it forces the user to alter his study to accommodate the software, much like (Epic and SAP) force the user to change business processes. Although, in industry, they likely use more comprehensive observation datasets that make this a moot point. We were limited to a single Payor dataset skewed to a particular segment of the population and considering its CMS, to the local population. But that's OK for this course.

Thank you.

# Next semester…

- We'll be teaching the OHDSI block again
  - Live class (come give a lecture at Georgia Tech!)
- Will expand the didactics to address some of the rough patches from this semester
- Maintain cloud-based Atlas but set up nodes for smaller units of the class (eg A-D, E-G, etc)
- Nuke the whole stack after the Labs in order to start fresh with Atlas, WebAPI, OHDSI DB
- Remove Atlas security

# Conclusion

- Should OHDSI be easy to use for all?
  - No, OHDSI is a scientific platform for scientists to do research
- BUT
  - It was challenging for even a couple of scientists (me and Jamie) to debug many of the issues found
  - As we look to deploy OHDSI environments at major scientific organizations (eg FDA, CDC, AMCs, pharma, etc), experiencing errors related to design or scale of users will set back adoption

# Massive Thanks

- James Wiggins (AWS)
- Jamie Weaver (Janssen R&D)
- …and all the awesome people who have built the many tools that I now have the luxury to gripe about.  I’m on the shoulders of giants.

# Questions?