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
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
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.

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

Power

Data source: Syrupf

Analysis: [follow-up] Basic Analysis

<table>
<thead>
<tr>
<th>Target subjects</th>
<th>Comparator subjects</th>
<th>Target years</th>
<th>Comparator years</th>
<th>Target events</th>
<th>Comparator events</th>
<th>Target IR (per 1,000 PY)</th>
<th>Comparator IR (per 1,000 PY)</th>
<th>MDRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,604</td>
<td>3,604</td>
<td>6,356</td>
<td>6,062</td>
<td>1,469</td>
<td>1,257</td>
<td>231.10</td>
<td>207.34</td>
<td>1.11</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Min</th>
<th>P10</th>
<th>P25</th>
<th>P50</th>
<th>P75</th>
<th>P90</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>1</td>
<td>120</td>
<td>391</td>
<td>733</td>
<td>977</td>
<td>1,005</td>
<td>1,088</td>
</tr>
<tr>
<td>Comparator</td>
<td>1</td>
<td>175</td>
<td>423</td>
<td>648</td>
<td>924</td>
<td>955</td>
<td>1,082</td>
</tr>
</tbody>
</table>

Survival probability

Number at risk:
- [Lab II] Hyperlipidemia Used Statin
  - 3,604
- [Lab II] Hyperlipidemia Never Used Statin
  - 3,070

Number at risk in days:
- 0: 3,604
- 250: 2,971
- 500: 2,441
- 750: 1,741

Number of deaths in days:
- 0: 5
- 250: 8
- 500: 13
- 750: 18
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
- OHDSI on AWS sets up automatically same credentials for Atlas and RStudio
So how did it go?
For Reference
Atlas Jobs on ohdsi.org

<table>
<thead>
<tr>
<th>#</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>02/26/2015 1:50 PM</td>
<td>02/26/2015 1:50 PM</td>
</tr>
<tr>
<td>19</td>
<td>02/26/2015 1:50 PM</td>
<td>02/26/2015 1:50 PM</td>
</tr>
<tr>
<td>18</td>
<td>02/25/2015 2:35 PM</td>
<td>02/25/2015 2:35 PM</td>
</tr>
<tr>
<td>17</td>
<td>COMPLETED</td>
<td>02/25/2015 2:34 PM</td>
</tr>
<tr>
<td>16</td>
<td>COMPLETED</td>
<td>02/25/2015 2:34 PM</td>
</tr>
<tr>
<td>15</td>
<td>COMPLETED</td>
<td>02/25/2015 12:43 PM</td>
</tr>
<tr>
<td>14</td>
<td>COMPLETED</td>
<td>02/25/2015 12:43 PM</td>
</tr>
<tr>
<td>13</td>
<td>COMPLETED</td>
<td>02/25/2015 11:56 AM</td>
</tr>
<tr>
<td>12</td>
<td>COMPLETED</td>
<td>02/25/2015 11:56 AM</td>
</tr>
<tr>
<td>11</td>
<td>COMPLETED</td>
<td>02/25/2015 10:16 AM</td>
</tr>
<tr>
<td>10</td>
<td>COMPLETED</td>
<td>02/25/2015 10:16 AM</td>
</tr>
<tr>
<td>9</td>
<td>COMPLETED</td>
<td>02/25/2015 10:12 AM</td>
</tr>
<tr>
<td>8</td>
<td>COMPLETED</td>
<td>02/25/2015 10:12 AM</td>
</tr>
<tr>
<td>7</td>
<td>COMPLETED</td>
<td>02/25/2015 10:11 AM</td>
</tr>
<tr>
<td>6</td>
<td>COMPLETED</td>
<td>02/25/2015 10:11 AM</td>
</tr>
<tr>
<td>5</td>
<td>COMPLETED</td>
<td>02/25/2015 10:08 AM</td>
</tr>
<tr>
<td>4</td>
<td>COMPLETED</td>
<td>02/25/2015 10:08 AM</td>
</tr>
<tr>
<td>3</td>
<td>COMPLETED</td>
<td>02/24/2015 8:30 PM</td>
</tr>
<tr>
<td>2</td>
<td>COMPLETED</td>
<td>02/24/2015 8:30 PM</td>
</tr>
<tr>
<td>1</td>
<td>COMPLETED</td>
<td>02/24/2015 8:20 PM</td>
</tr>
</tbody>
</table>

22,495 entries
02/24/2015 8:20 PM

As of 10/14/2019
Atlas Jobs on GT OHDSI

As of 10/14/2019

39,772 entries

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
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

**Benign essential hypertension**

- **Id**: 44834715
- **Code**: 401

**SNOMED (1)**
- **Id**: 44833566
  - **Code**: 401
  - **Name**: Essential hypertension

**ICD9CM (1)**
- **Id**: 312648
  - **Code**: 1201005
  - **Name**: Benign essential hypertension

**SNOMED (1)**
- **Id**: 314423
  - **Name**: Benign essential hypertension complicating pregnancy, childbirth and the puerperium - not delivered
  - **Person Count**: 27
  - **Prevalence**: 0.03%
  - **Length of Stay**: 0.20

- **Id**: 321080
  - **Name**: Hypertension complicating pregnancy, childbirth and the puerperium
  - **Person Count**: 21
  - **Prevalence**: 0.02%
  - **Length of Stay**: 0.00

- **Id**: 192679
  - **Name**: Renal disease in pregnancy AND/OR puerperium without hypertension
  - **Person Count**: 13
  - **Prevalence**: 0.01%
  - **Length of Stay**: 0.20
Had to search a level up to find

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

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.
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:

<table>
<thead>
<tr>
<th>a drug era of</th>
<th>[jyoo309] Ritalin</th>
</tr>
</thead>
</table>

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

Limit initial events to: earliest event \(\downarrow\) 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
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.
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.
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
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

```r
Exporting diagnostics
- covariate_balance table
  |============================================================
  | 100%
- preference_score_dist table
  |============================================================
  | 100%
Error in `colnames<-`(`*tmp*`, value = SqlRender::camelCaseToSnakeCase(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.

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

Before submitting your RStudio Errors Thread:

- Make sure you have lots of patients. If not:
  - Stop and follow-up
  - Rebuild your concept sets
  - If still 0 patients:

2) The other problem was described by Dr. Duke in @987 as "You have enough patients but the backgrounds for your patients are too divergent for the model 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...
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

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...

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 initial events to:

having all ▼ of the following criteria:
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
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
What was the DB load?

![Database Connections Chart]

- Tues
- Weds
- Thurs
- Fri
- Sat
- Sun
- HW Due!!
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

<table>
<thead>
<tr>
<th>Type of Job</th>
<th>Proportion of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort Generation</td>
<td>81.07%</td>
</tr>
<tr>
<td>Incidence Rate</td>
<td>12.04%</td>
</tr>
<tr>
<td>Characterization</td>
<td>5.30%</td>
</tr>
<tr>
<td>Other (eg cache)</td>
<td>1.59%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Job</th>
<th>COMPLETED</th>
<th>FAILED</th>
<th>STARTING</th>
<th>STOPPED</th>
<th>STOPPING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort</td>
<td>93.62%</td>
<td>1.84%</td>
<td>4.02%</td>
<td>0.49%</td>
<td>0.02%</td>
</tr>
<tr>
<td>IR</td>
<td>86.31%</td>
<td>3.50%</td>
<td>4.62%</td>
<td>5.49%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Characterization</td>
<td>78.51%</td>
<td>18.48%</td>
<td>0.00%</td>
<td>3.01%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Other (eg cache)</td>
<td>84.30%</td>
<td>11.13%</td>
<td>3.96%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Overall</td>
<td>91.79%</td>
<td>3.07%</td>
<td>3.88%</td>
<td>1.22%</td>
<td>0.02%</td>
</tr>
</tbody>
</table>
Atlas Job Performance

- 74% of students experienced at least one failed job (range 1 to 118 failures per student)
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. (Show Voters)
- **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 (and I) took some heat from the students

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!
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.
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.

ODHSI 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?