CVmedLab Manual

12/3/22

Table of contents

1 Welcome! 5

2 Introduction 6
- 2.1 Meetings . . . 6
  - 2.1.1 Lab meetings . . . 6
  - 2.1.2 1-on-1 meetings . . . 7
- 2.2 How we give feedback . . . 7
- 2.3 How we share things (and send them to Dr. Smith) . . . 8
  - 2.3.1 Code . . . 8
  - 2.3.2 Documents/Writing . . . 8
- 2.4 Shared lab resources . . . 9
- 2.5 References . . . 9

3 Lab culture and philosophy 10
- 3.1 Ask for help, and share your learning . . . 10
- 3.2 Make yourself available . . . 11
- 3.3 Come prepared and be engaged . . . 11
- 3.4 Celebrate accomplishments (yours & others') . . . 11
- 3.5 Sustain a positive, safe learning environment . . . 11
- 3.6 Have an interdisciplinary (open) mindset . . . 12
- 3.7 Be mindful and aware of your own biases . . . 12
- 3.8 Plan with intention, and follow through . . . 12
- 3.9 Foster inclusivity within our group and greater community . . . 12
- 3.10 Promote and sustain healthy work-life integration . . . 13
- 3.11 Practice radical candor . . . 13
- 3.12 Acknowledge and give credit . . . 13

4 Code of Conduct 14
- 4.1 Short version . . . 14
- 4.2 Longer version . . . 14
  - 4.2.1 Reporting . . . 16
  - 4.2.2 Consequences . . . 17
  - 4.2.3 Additional resources . . . 17
- 4.3 License and attribution . . . 17

5 Onboarding **18**
5.1 COP/POP Web resources . . . 18
5.2 Individual Development/Mentoring plans . . . 18
5.3 Facilities . . . 18
5.3.1 Office space . . . 18
5.3.2 Building & room access . . . 19
5.3.3 Parking and transportation: . . . 19

6 Expectations **20**

7 Offboarding **21**

8 Funding **22**

9 Communication **23**

10 Academics **24**

11 Courses **25**

12 Computing **26**

13 Data Resources **27**
13.1 Electronic Health Record (EHR) data . . . 27
13.1.1 OneFlorida+ . . . 27
13.1.2 UFHealth data . . . 27
13.2 Claims data . . . 28
13.2.1 CMS data (Medicare, Medicaid) . . . 28
13.2.2 Marketscan . . . 29
13.3 Others . . . 29
13.3.1 Clinical trial/prospective cohort data . . . 29
13.3.2 Publicly-Available datasets . . . 30

14 Coding Resources **32**
14.1 SAS . . . 32
14.1.1 Books . . . 33
14.1.2 Useful online articles/links/blogs . . . 33
14.1.3 Macros . . . 34
14.2 R . . . 34
14.2.1 Base R . . . 34

15 You need base R **34**
15.0.1 R-Studio and extensions . . . 35
15.0.2 R packages . . . 35

16 Installing Packages **36**

17 Production vs. Development packages **36**
17.0.1 Useful R resources 38
17.1 Git/Github 38
References 38

1 Welcome!

The focus of our work centers around studying effectiveness and safety of cardiovascular medications as well as the cardiovascular effects of other (non-cardiovascular) medications, with a focus on observational data methods and, to a lesser extent, clinical trials. More about our group's research activity can be found on our website.

This lab manual is intended to provide an overview for lab members and others about how we work and our expectations for our team. It is also a space to document institutional knowledge about procedures and available resources. If you have suggestions for additions or changes, please contact Dr. Smith or, if you're already github-savvy, make a pull request .

The spark for this book comes from a similar lab manual I found for The Fay Lab, which itself stemmed from the Openscapes Champions program.

This lab is licensed under a Creative Commons Attribution 4.0 International License.

2 Introduction

We do rigorous quantitative and epidemiologic science to support decision-making for cardiovascular disease treatment and prevention. To the extent possible, we conduct our work using Open Data Science principles, emphasizing scientific excellence (*not perfection*) that is transparent, reproducible, collaborative, and ethical. We aim to make our methods and results available and support ongoing learning.

Our quantitative work is based on sound design principles supported by statistical thinking, using evidence-based approaches to compare among alternatives for study designs and analytic options.

See the next chapter for more detail on our lab culture and philosophy. We are motivated heavily by the following two papers - which provide a blueprint for how we think about the way we do our work:

- Our path to better science in less time using open data science tools (Lowndes et al. 2017)
- Good enough practices in scientific computing (Wilson et al. 2017)

[To come: recommended reading list]

2.1 Meetings

The lab has two types of meetings: **lab meetings** and **1-on-1s** with Dr. Smith

2.1.1 Lab meetings

- **When we meet**: For 1 hour, every other week. The specific day/time will often change from semester to semester, depending on lab members schedules, but we aim to find a time that works for everyone.
- **How we meet**: Lab meetings are in person supplemented with Zoom for those unable to make it to campus. Students, in particular, should try to attend these meetings in person as much as possible.

- **What we discuss**: Determined by the lab members (and occasionally trumped by Dr. Smith to discuss pressing issues), but generally, these are research-related discussions. Most often, ongoing work in the lab, including, e.g., presenting draft study designs or preliminary analyses for feedback, discussing specific problems with an ongoing project and how we can overcome these. Sometimes topics may not be specific to a particular research project, but instead related concepts. For example, how to respond to reviewer comments, an overview of a new dataset being introduced to the lab, or a tutorial on creating a nice visualization.
- **Who decides what we discuss**: The lab members. One team member designated by Dr. Smith (typically senior GS or post-doc) coordinates the discussion topics for each lab meeting, and students should send discussion topics to them, or add to the agenda directly that is maintained on the General channel in CVmedLab Teams.
- **What you should expect**: You should expect to present fairly regularly, both to keep the lab updated on what you're doing, but also because nothing in our world goes perfectly smoothly. If you're not having problems you need to work through, the rest of us probably have concerns. You should also expect to contribute to discussions - we all have a unique background and set of experiences that can contribute meaningful insight to the discussion. Sometimes the ideas we think are the littlest (or possibly even worst) are in fact the most helpful.

2.1.2 1-on-1 meetings

- **When we meet**: For 1 hour, every other week (on the off-week from the whole lab meeting). The specific day/time will often change from semester to semester, depending on your schedule.
- **How we meet**: Dr. Smith prefers in person, but Zoom is also acceptable, if needed.
- **What we discuss**: Determined completely by you. This is your opportunity to discuss what is most pressing in your opinion: ongoing research, your IDP, school/class issues, any highlights or difficulties in the past couple of weeks, or just shoot the breeze. If there's something more complicated that you want Dr. Smith to know about/review in advance, make sure he gets this before the meeting (see below for how to get this to him).

2.2 How we give feedback

Feedback, both giving and receiving it, is an important aspect of our lab. Most of the feedback we give and receive is during lab meetings, on research products (e.g., abstracts, posters, papers), and when giving or attending seminars. We expect feedback to be supportive but constructive.

This resource from UBC does a really great job of outlining the main points of how to give and receive feedback.

2.3 How we share things (and send them to Dr. Smith)

It is useful to have standard ways of sharing things. These don't have to be followed absolutely, but should guide most of what we're sharing and make things easier on the team. When sending material to someone, always make sure to describe what you are sending and try to make it as easy as possible for them to help you.

Taking a project-based approach to organizing your work makes it easier to share and solicit feedback from others, as things tend to be self-contained. Try to keep only 1 working instance of material, and use some form of version control to facilitate this (see recommendations in Wilson et. al paper linked above).

Project management tools in Github are a good way to record and document questions on analyses, particularly if you're working in R and/or with analytic datasets that can be posted publicly. Use 'Issues' on github repositories for project-related tasks and problems. Unfortunately, that doesn't work for us in many cases due to data privacy/DUA issues with publicly posting patient-level data. Alternatively, make use of a Teams/Sharepoint (or, less preferably, Dropbox) for each project to record this history, much like you would a lab notebook.

2.3.1 Code

Code can be shared in the virtual machines, or alternatively through Teams, Dropbox, or Github repositories. For specific questions on problems, please try to create a reprex (minimal **rep**roducible **ex**ample). Ensure that others can run and interact with the material being shared.

2.3.2 Documents/Writing

Manuscripts and similar text documents can be created/shared in Quarto/R Markdown, if you're particularly motivated, or alternatively, in Dropbox or Teams. Quarto/R Markdown offer the advantage of being easily/quickly reproducible any time the underlying data change (e.g., w/o having to retype all the particular data points, re-do Tables, etc..). Dropbox allows for collaborative writing, and has the advantage of there only ever being one version (as opposed to files that are sent around via email). Teams/Sharepoint is also acceptable. Email is the least preferred approach for within-team collaboration, and should really be reserved (if needed) for getting comments from collaborators outside of our team. Word documents should always have your last name as the first part of the file name (please no "mythesis.doc").

We maintain a lab Teams sharepoint folder (accessible via Teams) for lab work, presentations, etc. Please make use of these so that others in the lab can make fair use of our work. Final publications will also be linked on our website and advertised on Twitter by the @UFCoDES and @CVmedLab accounts.

2.4 Shared lab resources

Where to find shared resources:

- Teams/sharepoint: You will be given access to Teams during onboarding; take a look at the General channel, which contains the lab meeting agenda/topics. You can create project-specific channels for your own projects, and add relevant lab members. Teams also offers access to sharepoint for the team, where you can store files.
- GitHub: CVmedLab is the shared GitHub account for the lab
- lab manual: This repository contains the lab manual, see section **??** for other useful resources
- Website: Take a look at the lab website, most of its information is duplicated in one of the above resources: http://www.cvmedlab.org/

2.5 References

3 Lab culture and philosophy

telos
the purpose, end, or goal of an entity

Our *telos* is truth. That means we pursue excellence in science - with the goal of finding truth foremost. But, we hold a lot of additional secondary goals and, importantly, we approach our science with a common denominator of kindness. As part of our core lab culture, we (in no particular order):

- **Ask for help, and share our learning**
- **Make ourselves available**
- **Come prepared and engage**
- **Celebrate accomplishments (ours & those of others)**
- **Sustain a positive, safe learning environment**
- **Have an interdisciplinary (open) mindset**
- **Are mindful of our own biases**
- **Plan with intention, and follow through**
- **Foster inclusivity within our group and greater community**
- **Promote and sustain healthy work-life integration**
- **Practice radical candor**
- **Acknowledge and give credit to other lab member's work**

Though we do not expect incoming members to memorize everything, we do expect members to be aware of the group's values.

3.1 Ask for help, and share your learning

We are all learners, and most of our learning is done from each other. It is inefficient to struggle through problems alone. Take advantage of the team, as well as the larger GS program. Ask for and give assistance with appropriate cognizance of the value of your time and the time of the person you are asking. You are not the first or last person to encounter a problem. When you identify a problem add an issue to the lab issues repository, and update it with solutions when it is resolved. Also consider writing a tutorial/blog post for inclusion in our lab's shared resources, and share with the larger community by tweeting, leading and sharing at a lab meeting, or running a workshop (Quantfish woRkshop group, SouthCoast useR group, etc).

3.2 Make yourself available

Be responsive to communication, and make time for things that address longer term goals, even when busy. For example, do not skip on things like attending seminars just because you have a big deadline looming (see note below on planning and organization). Note that being available does not mean that you are available 24/7 - this is not expected. Because as a group we value the role of collaboration and interaction in improving our work, it is expected that you be available in the office/campus during normal business hours for some time during the week (see section on attendance expectations).

3.3 Come prepared and be engaged

Value your time. Be present during lab and individual meetings, and come to your work ready to do your work. Contribute and participate in planning and lab discussions. When it is your turn to run a meeting, come with an agenda and be prepared with questions. Aim to view meetings as events that contribute to your work and productivity, rather than taking away from them.

3.4 Celebrate accomplishments (yours & others')

You and your colleagues work hard. Things don't always go exactly as planned. Be supportive and proud of yourself and your peers when you accomplish things. We are not competing with each other - someone else's success does not mean your failure. Share your accomplishments with others!

3.5 Sustain a positive, safe learning environment

We're here to learn and grow as scientists. Everyone learns something for the first time at some time, and people learn in different ways. Expressing that you don't know something is OK - good, even! - this is a University after all. Nevertheless, we understand that this can make you feel vulnerable. We strive to maintain a culture that allows for and encourages this vulnerability. Community members should not be disparaged for not knowing things, and in addition, should not be disparaged for knowing things or wanting to learn. That's precisely why we're all here.

3.6 Have an interdisciplinary (open) mindset

It's quite rare that any one person has the necessary clinical and methodologic expertise to go at a project alone. It's even rare that a single person has the necessary experience/knowledge to sufficiently cover one of those arms. We collaborate, both within the CVmedLab, as well as outside it, because we most often work on problems that span multiple disciplines. Co-creation of knowledge requires transdisciplinary approaches that can result in solutions that would not be possible with siloing. You will be collaborating with others who have different types of expertise, values, and terminology. Trust the expertise of others and actively seek feedback recognizing the importance of specialization.

3.7 Be mindful and aware of your own biases

We all have biases that are inherent and can not be removed, but we can still work on both being less biased, and more aware of bias in ourselves and others. Periodically check in on your biases.

3.8 Plan with intention, and follow through

Be organized and adaptable. Things don't always go as planned and that's OK. Planning can help you adapt when they don't (see Come Prepared). Find a program/project management approach that works for you (see "How we work"); being organized can reduce stress immensely and help you progress with your goals.

3.9 Foster inclusivity within our group and greater community

Part of our lab culture is that we are good citizens of our community, we take on leadership roles, when feasible, within UFCOP and the University, we are supportive of others in our community during their milestones, we actively participate in COP/POP events (e.g., seminar, college talks, etc) and perform outreach. Work with Dr. Smith when crafting your individual mentoring/development plans (see Chapter on Onboarding) to identify what you want to (re-)aim your efforts at.

3.10 Promote and sustain healthy work-life integration

Our scientific research is not the only important thing in our lives, and publishing research is not the only mechanism by which to provide science and support our communities. We recognize the importance of our other commitments in keeping us healthy (mentally and physically) and bring our whole selves to our efforts. Try not to normalize overwork or being busy as achievement or status. Plan downtime for yourself to recuperate.

3.11 Practice radical candor

We care personally while also challenging directly. Be honest when communicating, accept critical (but kind) feedback, and give the same to others. View relationships within the group as collaborative rather than evaluative. Don't take constructive criticism personally. No one is critiquing you, the person; we're focused on making (and communicating) the science optimally.

3.12 Acknowledge and give credit

Working as part of a team, we will almost always be building on work done by others, receive assistance with work (see "Ask for help"), and using others' words, code, philosophy, or content. Include acknowledgement and give credit for those contributions, in all forms of communication. We share content and code within the group with this expectation. One easy approach is to include hyperlinks to the work of others or their social media in your work. This also helps to amplify their work (and voice) as well as yours.

4 Code of Conduct

> **Code of Conduct** a set of basic ground rules that participants (in our lab) are expected to follow.

The goal of having a code of conduct is to create an open and inclusive space for our work that helps us achieve our collective goals. Along with our lab culture/philosophy, it also provides a benchmark for self-evaluation and helps better define our identity as a community.

We expect all lab members to adhere to the policies and guidelines outlined here.

Additional information and resources can be found in the UF Student Conduct Code and Honor Code (fair warning, it's long, but does have some important stuff in there).

(https://sccr.dso.ufl.edu/process/student-conduct-code/).

4.1 Short version

The CVmedLab is dedicated to providing a harassment-free experience for everyone, regardless of gender, gender identity and expression, sexual orientation, disability, physical appearance, body size, age, race, or religion. We do not tolerate harassment of participants in any form.

This code of conduct applies to all lab spaces and interactions, including group and individual meetings (face to face and remote), workshops, social events, email correspondence, and web channels and code repositories, both online and off. Anyone who violates this code of conduct may be sanctioned and referred to the university's academic policies.

4.2 Longer version

The Fay lab is dedicated to providing a harassment-free experience for everyone. We do not tolerate harassment of group members or others in our larger communities in any form. Bullying is unwanted, aggressive behavior that involves a real or perceived power imbalance. The University of Massachusetts Dartmouth's Equal Opportunity and Anti-Harassment policies, complaint procedures and form provides definitions and examples of harassment and sexual harassment.

This code of conduct applies to all Fay lab spaces, including group and individual meetings (face to face and remote), workshops, social get togethers, email correspondence, and web channels and code repositories, both online and off. Anyone who violates this code of conduct may be sanctioned and referred to the School's and university's academic policies.

Some Fay lab spaces may have additional rules in place, which will be made clearly available to participants. Participants are responsible for knowing and abiding by these rules.

Harassment includes:

- Offensive comments related to gender, gender identity and expression, sexual orientation, disability, mental illness, neuro(a)typicality, physical appearance, body size, age, race, or religion.

- Unwelcome comments regarding a person's lifestyle choices and practices, including those related to food, health, parenting, drugs, and employment.
- Deliberate misgendering or use of 'dead' or rejected names.
- Gratuitous or off-topic sexual images or behaviour in spaces where they're not appropriate.
- Physical contact and simulated physical contact (eg, textual descriptions like "*hug*" or "*backrub*") without consent or after a request to stop.
- Threats of violence.
- Incitement of violence towards any individual, including encouraging a person to commit suicide or to engage in self-harm.
- Deliberate intimidation.
- Stalking or following.
- Harassing photography or recording, including logging online activity for harassment purposes.
- Sustained disruption of discussion.
- Unwelcome sexual attention.
- Pattern of inappropriate social contact, such as requesting/assuming inappropriate levels of intimacy with others
- Continued one-on-one communication after requests to cease.
- Deliberate "outing" of any aspect of a person's identity without their consent except as necessary to protect vulnerable people from intentional abuse.
- Publication of non-harassing private communication.

The Fay lab prioritizes marginalized people's safety over privileged people's comfort. The PI (Gavin Fay) reserves the right not to act on complaints regarding:

- 'Reverse' -isms, including 'reverse racism,' 'reverse sexism,' and 'cisphobia'
- Reasonable communication of boundaries, such as "leave me alone," "go away," or "I'm not discussing this with you."
- Communicating in a 'tone' you don't find congenial
- Criticizing racist, sexist, cissexist, or otherwise oppressive behavior or assumptions

4.2.1 Reporting

If you are being harassed by a member of the Fay lab, notice that someone else is being harassed, or have any other concerns, please contact the PI, Dr. Gavin Fay, at gfay@umassd.edu. If you do not wish to contact Dr. Fay, please contact Department Chair Dr. Steve Cadrin scadrin@umassd.edu and/or SMAST Assistant Dean Mike Marino mmarino@umassd.edu. If the person who is harassing you is on the team, they will recuse themselves from handling your incident. We will respond as promptly as we can.

Lab members who believe they have been subjected to any kind of discrimination that conflicts with the University of Massachusetts' policies and/or the laws of the Commonwealth of Massachusetts' should seek assistance from a supervisor or an HR representative. Information about the University of Massachusetts Dartmouth's Equal Opportunity and Anti-Harassment policies, complaint procedures and form.

Confidentiality: Please remember that by way of his position at the university, Dr. Fay is a mandated reporter under Title IX. This means that he is not allowed to keep matters falling under Title IX confidential, and is required to disclose these incidents to the administration. You are welcome to discuss matters with Dr. Fay, but please keep this in mind when doing so. Gavin will do his best to remind you of his responsibilities at the start of conversations anticipated to relate to these topics. If you would prefer to discuss incidents of harassment that fall under Title IX with a confidential source, you can contact David Gomes, UMass Dartmouth Director of Diversity and Inclusion.

This code of conduct applies to Fay lab spaces, but if you are being harassed by a member of the Fay lab or another member of the SMAST community outside our spaces, we still want to know about it. We will take all good-faith reports of harassment by Fay lab members seriously. This includes harassment outside our spaces and harassment that took place at any point in time. The abuse team reserves the right to exclude people from the Fay lab based on their past behavior, including behavior outside Fay lab spaces and behavior towards people who are not in the Fay lab. The University Counseling Center provides a range of services to help students develop improved coping skills to address emotional, interpersonal and academic concerns.

To protect our team members from abuse and burnout, we reserve the right to reject any report we believe to have been made in bad faith. Reports intended to silence legitimate criticism may be deleted without response.

The above caveats noted, we will respect confidentiality requests for the purpose of protecting victims of abuse when possible. At our discretion, we may publicly name a person about whom we've received harassment complaints, or privately warn third parties about them, if we believe that doing so will increase the safety of the Fay lab members or the general public. We will not name harassment victims without their affirmative consent.

4.2.2 Consequences

Participants asked to stop any harassing behavior are expected to comply immediately.

If a participant engages in harassing behavior, Dr. Fay may take any action they deem appropriate, which includes referral to the Department Chair and Dean's Office under the SMAST policy, and also including expulsion from all Fay lab spaces.

4.2.3 Additional resources

Other University of Massachusetts Board of Trustees and University documents:

- Policy Against Intolerance
- Principles of Employee Conduct
- Resolution in Support of Pluralism
- Non-Discrimination and Harassment Policy
- Statement of Affirmative Action and Equal Opportunity
- Statement on Cultural Diversity and Inclusion
- Statement on Gender Discrimination
- Bully, Harassment and Violence
- Student Conduct Policies and Procedures
- Sexual Harassment Policy
- Title IX and Sexual Assault/Harassment

4.3 License and attribution

This anti-harassment policy is based on the example policy created by the Geek Feminism community. It has also been written to fall within the SMAST Code of Conduct and Diversity Statement, and has been reviewed by the SMAST Dean's office and UMass Dartmouth Departments of Human Resources and Office of Diversity Equity and Inclusion.

5 Onboarding

Welcome to the CV med Lab! First, we are excited that you have decided to join our team! We hope that these onboarding resources, guidelines, and tips will make your transition to the team, Department, College, and University seamless and enjoyable.

5.1 COP/POP Web resources

POP/COP related resources, forms, policies, procedures, calendars, as generally housed on, or linked from the respective websites for POP and COP. A couple of specific COP sites will probably be more helpful to you, including the COP research office and the COP Grad Education office.

5.2 Individual Development/Mentoring plans

Within the first few weeks of joining the program, you will learn about the IDP requirements for the COP graduate program. You should work with Dr. Smith to develop a plan outlining your short, medium, and long term goals early in year 1, and these will need to be revisited at least annually. More on individual developing plans can be found in Chapter @ref(expectations).

5.3 Facilities

5.3.1 Office space

The CVmedLab has no devoted physical lab. Most of our work is done in virtual space. POP GS students have access to shared office space in HPNP room 2###; some desks are assigned and others are reserved for temporary use. Assigned desks are coordinated by the POP graduate student representative(s). Post-docs and analysts typically have individual or shared office space on either the 2nd or 3rd floor of HPNP. Dr. Smith's office is HPNP 3316.

Note that we will soon be moving to the new Data Science building, in which we'll have more space, including devoted work stations for all GS students.

Lab meetings are currently held in shared conference rooms (usually HPNP 2306 or 2309).

5.3.2 Building & room access

Access to the HPNP building is via GatorOne. Office space is keyed, and you should get a key for accessing the shared POP GS office. The same key allows access to the shared conference rooms.

5.3.3 Parking and transportation:

Parking on campus is less than optimal, as spaces are limited, fairly expensive, and particularly for students, not very close to HPNP. Most students bus, bike, or walk to campus. Full details for campus parking policies and procedures are on the TAPS webpage.

6 Expectations

More to come...

7 Offboarding

Still working on this.

8 Funding

More to come...

9 Communication

More to come...

10 Academics

Still need to do this...

11 Courses

Need to do this...

12 Computing

To do.

13 Data Resources

POP/CoDES have a wealth of data resources that are generally refreshed/updated every 1-2 years, and available to the graduate program, including you. Our lab uses primarily OneFlorida+ data (which includes FL Medicaid), Marketscan, and Medicare. Read on below for info on each. You can access to any/all of the above for your own independent research projects or thesis work, though some have fees attached which you'll need to discuss with Dr. Smith.

Most of these data are housed on ResVault virtual machines or the POP high-performance server. See Chapter 12 for more information.

13.1 Electronic Health Record (EHR) data

13.1.1 OneFlorida+

OneFlorida+ is one of 11 clinical research networks (CRNs) that comprise the Patient-Centered Ourcomes Research network (PCORnet). Quite a bit of our work is done on OneFlorida+ data, or data from OneFlorida+ and other PCORnet CRNs. The good news is they all adhere to the PCORnet common data model (scroll to the bottom of the page), which you will need to familiarize yourself with if you're working on OneFl+/PCORnet data. In particular, you'll need to familiarize yourself with the relevant tables and variables. If you're unsure whether you'll be working with OneFl+/PCORnet data, ask Dr. Smith.

OneFlorida+ includes EHR data from health system partners across Florida (UF, University of Miami, University of South Florida, Orlando Health, Florida Hospital [Orlando], Tallahassee Memorial, and others), as well as University of Alabama-Birmingham (UAB) and Emory University. In addition, it contains Florida Medicaid claims data. Data are generally available from ~2012 onward. As you'll see on reviewing the common data model, the available data are generally structured data (rather than unstructured clinical texts like provider notes, imaging reports, etc..).

13.1.2 UFHealth data

The UF integrated data repository is a database of UFHealth EHR data, including both structured and unstructured data. The IDR has an i2b2 implementation which can be used for simple queries, i.e., to find counts of patients that meet certain criteria. The IDR i2b2

implementation can be found here (you'll need to register for an account here and you'll need to be on campus or on the HSC VPN to access i2b2).

We are currently in the process of linking UFHealth data with our Medicare claims data for patients in both data sources.

13.2 Claims data

Major claims data sources housed in POP/CoDES include CMS data and IBM Marketscan. Brief descriptions are below. Additional information can be found here.

13.2.1 CMS data (Medicare, Medicaid)

13.2.1.1 Medicaid

Medicaid Analytic eXtract (MAX) and T-MSIS Analytic Files (TAF) data contain claims for medical care and drug benefits received by beneficiaries with Medicaid insurance coverage, the state-run programs for low-income and categorically eligible individuals and families. CoDES has in-house MAX data for over >120 million beneficiaries residing in the 29 most populous states from 1999-2010 (AL, AR, CA, FL, GA, IA, ID, IL, IN, KS, KY, LA, MA, MN, MO, MS, NC, NE, NJ, NM, NY, OH, SC, TN, TX, VA, WA, WI, WV) and national data (all 50 states plus the district of Columbia) from 2011-2016.

Medicaid data has been linked to birth certificates from the Florida Department of Health (1999-2014), Texas Department of State Health Services (1999-2012) and New Jersey Department of Health (1999-2010). The entire national Medicaid data set includes validated mother-infant linkages.

13.2.1.2 Medicare

Medicare data include claims for inpatient, skilled care nursing facility, and hospice care (Part A) as well as outpatient care (Part B) and prescription drugs (Part D). CoDES center has a somewhat complicated sample of Medicare, due in part to our desire to link UFHealth EHR data with Medicare data.

Basically, the current sample includes the following:

- A 5% national Medicare sample (random sample of 5% of Medicare patients nationwide who meet the above criteria for parts A, B, and D) for the years 2011 through 2015 plus 1 million beneficiaries in FL sampled from individuals who reside in the UF Health catchment area (to ensure we could link most UFHealth patients)

- A 15% national Medicare beneficiaries plus the entire state of Florida for 2016-2018, totaling >8 million lives.

We are anticipating continuing to grow the data (additional years).

ResDAC contains excellent documentation of the Medicare files, variables, and availability from year-to-year. If you're going to use Medicare data, you'll need to get to know these data dictionaries.

13.2.2 Marketscan

The IBM Marketscan Commericial claims database includes 2005-2020 health insurance claims for inpatient, outpatient, and outpatient pharmacy encounters, as well as enrollment data from large employers and health plans across the United States who provide healthcare coverage for their employees, their spouses, and dependents. The current dataset includes >192 million lives.

The Medicare Supplemental data includes 2005-2020 enrollment records along with inpatient, outpatient, ancillary, and drug claims for >12.9 million retirees in the United States with Medicare supplemental coverage through privately-insured fee-for-service, point-of-service, or capitated health plans.

The Health Risk Assessment (HRA) data includes 2012-2018 self-reported biometric and health-related behavioral data obtained through surveys of employees of large US corporations and health plans. HRA is linked to medical, pharmacy, and enrollment data for these employees in the Commercial claims data (above) and used to examine the relationships between health behaviors/risk and health outcomes or medical expenditures. Linked data is available for about 5% of beneficiaries.

13.3 Others

There's much too much to make this a comprehensive list, but here are some additional data resources that are either publicly-available or available to us by virtue of collaborations within UF, and may be of interest to you/the lab for some of our work.

13.3.1 Clinical trial/prospective cohort data

- NHLBI BioLINCC - NHLBI-funded clinical trial and prospective cohort data

> **Note**
>
> We currently have access to SPRINT and ACCORD trial data - ask Dr. Smith if interested being added to the DUA for these trials

- INVEST trial - we have access to the INVEST trial data, which was a large international trial (22.5k individuals enrolled) comparing a calcium channel blocker vs. beta-blocker treatment strategy in patients aged 50 years with hypertension + coronary artery disease. Includes adjudicated cardivoascular events, as well as all-cause death data through at least 2015.
- WISE cohort - we have access to the Women's Ischemia Syndrome Evaluation (WISE) cohort, which was a multisite prospective cohort study of women with suspected myocardial ischemia.
- Women Take Heart - we have access to the Women Take Heart cohort, which was a Chicago-based prospective cohort study of ~8k women without cardiovascular disease, enrolled in 1992 and with death follow-up through at least 2008.
- WARRIOR trial - The Women's IschemiA TRial to Reduce Events In Non-ObstRuctive CAD is a multicenter, prospective, randomized, blinded outcome evaluation (PROBE design) of a pragmatic strategy of intensive medical therapy (incl. ACEI or ARB + statin) vs usual care in 4,422 symptomatic women with ischemia and no obstructive coronary artery disease (INOCA)

13.3.2 Publicly-Available datasets

The CDC curates a number of valuable datasets that are relatively easy to access and generally offer cleaned, curated datasets that are analysis-ready. Some common ones we use/see in our field include:

- National Health and Nutrition Examination Survey (NHANES) - a complex survey design that is completed every 2 years and allows for inference about what is happening across the non-instutitionalized U.S. population
- National Ambulatory Medical Care Survey (NAMCS) - Data provided by *providers* (not patients) about patient visits in a single week of the year; allows for inference about what is happening at outpatient visits in the U.S.
- Behavioral Risk Factor Surveillance System (BRFSS) - state-administered surveys, completed annually, and curated by the CDC
- And, lots of others from the National Center for Health Statistics

- Medical Expenditure Panel Survey (MEPS) - a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States; MEPS is the most complete source of data on the cost and use of health care and health insurance coverage in the U.S.
- FDA Adverse Event Reporting System (FAERS) - datasets containing Adverse Events reported to the FDA on drugs; there is a similar reporting system administered by DHHS for vaccines, called VAERS

14 Coding Resources

Learning to code for data wrangling and analyses will be a significant component of your education during the MS or PhD program. Some people will take these skills forward and continue using them in their career after graduate training, but even if you don't, use these skills extensively yourself as you move into your career, you will likely be overseeing people who do, and it's good to know general principles of coding, how problems arise in coding that are sometimes difficult to diagnosis, and how to overcome these problems, even if you're not the one directly coding your analyses for the rest of your career.

What will you learn in our program? Primarily SAS and R, but you're welcome to explore other platforms as well during your time here.

> **Note**
>
> Our program considers SAS the defacto data wrangling/analysis platform and that's what you'll use/be exposed to in most of the Departmentally-administered courses. It's also commonly used in courses administered by the Biostatistics Department, some of which are required for you during the program.
> That said, there's a lot to love about R, and some good reasons you might want to have this in your repertoire as well. For starters, it's open-source and free, enhanced frequently, and it has an excellent ecosystem of extensions (called "packages") that allow anyone (including you!) to add additional functionality for the R community. Perhaps most importantly, R creates markedly better publication-ready graphics than SAS does, and with less effort.

Other platforms, for example, SPSS and python, get some use in our department, but are not particularly widespread, though perhaps that will change (particularly for python) with the new AI initiatives at UF.

14.1 SAS

As noted above, SAS is the primary platform used in our program. We use SAS fairly extensively in our lab's work as well, in part because it's particularly good at working with massive datasets. SAS is available on all of our Virtual Machines/servers, and UF offers discounted annual licenses for SAS (as well as a free cloud-based SAS through UFApps) for enrolled

students. More info on individual licenses can be found here for students and here for faculty/postdocs/staff. Note that staff/postdocs should get their license through the Department by contacting Carl Henriksen.

Some folks like working in base SAS by itself. Others prefer SAS Enterprise Guide, which wraps around base R and provides some additional functionality. Try each, and see what you prefer.

One downside to SAS is it does not run natively on MacOS, so if you have a Mac, you'll need Parallels, VMware, or similar hardware virtualization to create a windows drive, if you want SAS on your own system.

14.1.1 Books

There are lots of good SAS books out there, but here's a couple you might find particularly useful. (* denotes texts Dr. Smith has electronic copies of and that can be 'checked out' within the lab.)

- The little SAS book, 6th ed. (you can access this one from campus or on the VPN - Dr. Smith also has a copy of 5th ed.*) (Delwiche and Slaughter 2019)
- Analysis of Observational Health Care Data using SAS* (Faries and Institute 2010)
- Survival analysis with SAS: A practical guide* (Allison 2010)

14.1.2 Useful online articles/links/blogs

- SAS Procedures by Name - This is a must-have bookmark to the official SAS documentation; you will use it often and it's quite helpful.
- UCLA Office of Advanced Research Computing tutorials - a good starting place for basics of running relatively simple analyses/data wrangling and interpreting.
- UF PHC 6052 course tutorials - you'll take this class relatively early in the program, but still a useful resource
- The DO loop - excellent and very productive blog by Rick Wicklin
- LexJansen.com - not a particularly user-friendly site, but contains tons of SAS-related papers. Your best bet is just googling your problem, but there's a good chance the top hits will be papers in PDF form on this site.

14.1.3 Macros

- Squeeze - shrink datasets by minimizing variable lengths to minimum necessary for the actual dataset
- [Magic Macro] - to add
- Basic Dataset Characterization
- OptionReset - reset default options, if you've somehow mangled yours
- ms_freezedata - a mini-SENTINEL program macro that creates subsets of patient-level data from a supplied patient id list
- [Table 1] - to add

14.2 R

14.2.1 Base R

You can download base R for free from C-RAN here. Make sure you select the correct file for your computer system (and chip, if using a Mac).

Danger

15 You need base R

You need to install base R, even if you will use R-Studio (recommended). Otherwise, R-Studio won't do you much good.

Installation should be straight-forward and easy, and you can use defaults. If needed, there are comprehensive instructions (HTML and PDF) available on C-RAN.

The R development team has some good, if somewhat dense, manuals available at C-RAN here. A good place to start is the Intro to R (HTML and PDF).

Tip

Base R is updated pretty frequently, but you won't be bugged about it. It's a good idea to check periodically to see if you are behind a few releases. The easiest way to update is to just re-install the new version, using the same process you did the first time around (download the compiled software and re-install; it will write over the old version).
Unlike python, the R development team takes great pains to not break things with any new releases, so it's rare that an update will cause you any problems with old code.

15.0.1 R-Studio and extensions

We recommend you use R-studio on top of R. Select the free desktop version here (*note: skip to step 2 since you will have hopefully already installed base R*). Again, make sure you download the correct file for your operating system and note that the website often assumes you run Windows - if you don't, scroll down a bit further to make sure you get the right file for your OS.

Again, installation should be straight-forward and easy, and you can use defaults.

There are some useful extensions for R-studio, but none are absolutely necessary:

- Quarto - useful for generating all sorts of R-markdown, including papers (yes, you could actually write your manuscript in Quarto), technical reports, websites, books (this lab manual was written with Quarto!); the beauty of R-markdown is you can weave together plain text and R code seamlessly into an output document (.html, .docx, .pdf, etc) - that means you can have all of your analysis code re-generate everything automatically any time you make a slight change to the cohort or underlying data.
  - One 'to-do' for the lab might be to make a manuscript template which would make writing routine parts of manuscripts considerably easier

15.0.2 R packages

Tip

16 Installing Packages

R packages are installed by typing `install.packages("package_name")` in the interactive window (or in a new script, which is then run). For example, `install.packages("tidyverse")`. You'll sometimes see 'dependencies' installed as well - these are other packages that are required to run the package you're installing. If you get questions about compiling from source, just select No unless you really need that slightly newer version.

Note

17 Production vs. Development packages

The below listed packages are all hosted on C-RAN and can be installed with the `install.packages()` function. You may come across packages in development that you want to install, or development versions of established packages that have not been pushed to C-RAN yet. These can usually be installed from, e.g., github, using the devtools package, with something like `devtools::install_github("github_user/package_name")` and typically the package will supply a similar instruction if you find the associated webpage or github page.

17.0.0.1 Data wrangling

The following packages are particularly useful for dealing with raw data as well as basic analyses:

- tidyverse - a suite of packages that make R considerably easier to learn for the new user (in our opinion); bonus points because they're supported by Posit (makers of R-Studio) and are constantly being improved, unlike some packages which eventually languish. We

suggest installing the entire tidyverse with `install.packages("tidyverse")` but you can also install components of the tidyverse individually

- data.table
- labelled
- []

17.0.0.2 Graphics & Tables

- ggplot2 - comes with tidyverse, so does not need separate installation, but will be your go-to for plotting much of what you'll want.
  - ggplot2 also has lots of very useful extension, which can be found here. Some we particularly like include patchwork, ggrepel, ggthemes, ggsci, ggdist, ggsignif, and survminer (there are >100!)
- plotly - interactive graphics
- gt
- gtExtras - extension for {gt}
- gtsummary - extension for {gt} for creating summary tables. **See Lab Docs on the CVmedLab website for an example using gtsummary.**

17.0.0.3 Interactive data presentation

- shiny - builds interactive web apps from R

17.0.0.4 Package development

Packages come in all shapes and sizes and don't have to be a set of fancy functions to be used by the R community. A common use of R packages is collating everything needed for an analysis/paper (data, notes, analytic code, +/- the paper itself). If you're going to create packages, the following are extremely helpful:

- devtools - simplifies common tasks in package development (also helpful in downloading non-C-RAN packages/versions, e.g., from github)
- usethis - automates repetitive tasks in package development

17.0.0.5 Analysis

Tons of packages in this space, but if you want to stick with the tidyverse, tidymodels is a good choice.

For regression modeling, Frank Harrell's rms package is good and well supported with a website.

17.0.0.6 Other Odds-and-Ends

- daggity - for creating DAGs and much more
-

17.0.1 Useful R resources

17.1 Git/Github

- Happy Git with R

References

Allison, Paul D. 2010. *Survival Analysis Using SAS: A Practical Guide*. 2. ed. Cary, NC: SAS Press.

Delwiche, Lora D., and Susan J. Slaughter. 2019. *The Little SAS Book: A Primer*. Sixth edition. Cary, NC: SAS Institute.

Faries, Douglas E., and SAS Institute, eds. 2010. *Analysis of Observational Health Care Data Using SAS*. Cary, North Carolina: SAS Publishing.

Lowndes, Julia S. Stewart, Benjamin D. Best, Courtney Scarborough, Jamie C. Afflerbach, Melanie R. Frazier, Casey C. O'Hara, Ning Jiang, and Benjamin S. Halpern. 2017. "Our Path to Better Science in Less Time Using Open Data Science Tools." *Nat Ecol Evol* 1 (6): 1–7. https://doi.org/10.1038/s41559-017-0160.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. "Good Enough Practices in Scientific Computing." *PLOS Computational Biology* 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.