PHPM 672/677 Final Project (Revised)

Due date: Submit in E-Campus by 11:59pm Tues 5/5

Presentation: Thursday 04/30 3-5pm

Milestones 1 &2 due date: E-Campus by 11:59pm 4/13 & 4/20

Submission. Submit on E-Campus by 11:59pm the day before the class they are due.

- Milestone 1: Due in One week
  - List of datasets & a list of variables you plan to use for the questions (does not have to be the final list). You must have at least 3 datasets.
  - List of questions you will answer & intro: Need to be sufficient. I will approve. If you are working in groups, each student must have some questions they are responsible for answering and writing the code to answer the question. Designate who is in charge of which questions
  - If working in groups: how the work will be split up (i.e. roles)
- Milestone 2: Due in Two weeks
  - A short write up on
    - what has been completed
    - and what is left to do
  - work in progress program (whatever you have. it does not need to be completed)
- Final Presentation: 4/30 3-5pm
  - Presentation (15-30 minutes. Will finalize after we know how many groups)
    - Must include a section on the data pipeline (which tables were combined how) and a "data flow chart" (# of observations over the data flow. See example below)
    - Must include a section on variables used (where did the variables come from and how where they calculated, if applicable)
- Final Submission: 5/5 midnight
  - All final program, log, and output (lst or html).
  - **[Only for PHPM672. Not required for PHPM677]** Write up: Not formal paper
    - (1) Hypothesis, (2) Describe Data & Method, (3) Questions & Answers

Recommended readings for this assignment

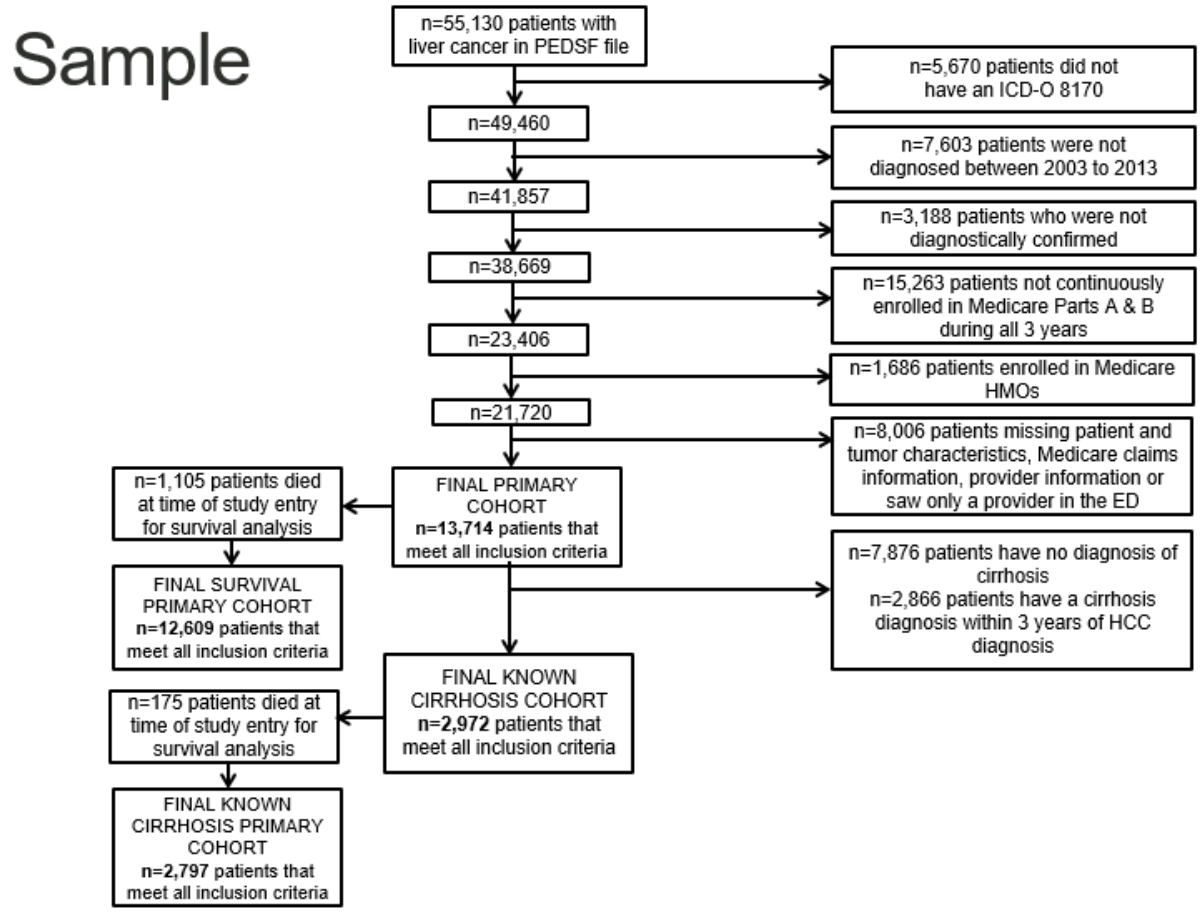
Below are some papers that use large secondary data for health services research. There are many others. You should read a few to understand what kinds of questions are addressed how. The general steps are as follows. For this class, you are focusing on steps 3 to 8. So do your best to define some questions that is doable in the given time for your team.

1. Identify a topic of interest
2. Conduct a literature search to understanding what is known and what are the next questions to ask (publishable). Also, you can learn what are potential sources of data
3. **Identify data source you will use to answer your question.**
   a. Most likely, you will have to adjust your research question to something that can be answered using the data you have access to
4. **Once data is confirmed, identify specific research questions you will answer using the data at hand**
   a. For each question, think of a hypothesis that the question would confirm or not
5. **Plan out the analytic data set you need to answer your questions**
6. **Plan out a data pipeline to construct the analytic data set from the raw data sources**
7. **Implement the data pipeline**
8. **Analyze the analytic data set you created to answer your questions (given the nature of this class, you are not expected to do anything fancy in the analytics.)**
9. Conclude on whether your hypothesis was met or not, and implications of your findings

1. Choi, D. T., Kum, H. C., Park, S., Ohsfeldt, R. L., Shen, Y., Parikh, N. D., & Singal, A. G. (2019). Hepatocellular carcinoma screening is associated with increased survival of patients with cirrhosis. *Clinical Gastroenterology and Hepatology*, *17*(5), 976-987.
2. Phillips, C., Truong, C., Kum, H.-C., Nwaiwu, O., and Ohsfeldt, R. (2017) A Population Study of Post-Acute Care for Children with Special Health Care Needs in Texas, 2011-2014. *Clinical Medicine Insights: Pediatrics*.
3. John Billings and Maria C. Raven Dispelling. An Urban Legend: Frequent Emergency Department Users Have Substantial Burden Of Disease. Health Affairs, 32, no.12 (2013):2099-2108
4. Yung, R. L., Chen, K., Abel, G. A., et al. (2011). Cancer disparities in the context of Medicaid insurance: a comparison of survival for acute myeloid leukemia and Hodgkin's lymphoma by Medicaid enrollment. The oncologist, 16(8), 1082-1091.
5. Bronstein J, Lomatsch C, Fletcher D, Wooten T, Lin TM, Nugent R, Lowery C. Issues and Biases in Matching Medicaid Pregnancy Episodes to Vital Records Data: The Arkansas Experience. Maternal and Child Health Journal, 2009;13(2):250-259

Data flow chart

Below is an example data flow chart for paper 1 (Choi et. al, 2019) above. Note that the figure below does not have the details of the full data pipeline depicting what tables were put together how to get n=55,130 patients that is the start of this data flow chart. That was described in the paper. For your presentation and paper, you will have to also include this information in addition to the data flow chart. Learning how to think about, understand, and communicate about this process is one of the main learning objectives of his course.



Guideline for assignment grading (Total of 32)

PHPM 672	PHPM 677

• Milestone 1 (Total 4)
• Milestone 2 (Total 4)
• Final Presentation (Total 24)
  ○ Program (final): 8 points
  ○ Presentation: 4 points
  ○ Write up (final): 12 points
  ▪ Hypothesis (2)
  ▪ Describe Data & Method (4): Make sure to describe how your datasets were combined including the any linkage relationships
  ▪ Questions & Answers (6)
	• Milestone 1 (Total 4)
	• Milestone 2 (Total 4)
	• Final Presentation (Total 24)
	  ○ Program (final): 12 points
	  ○ Presentation: 12 points
	  ▪ Hypothesis (2)
	  ▪ Describe Data & Method (4): Make sure to describe how your datasets were combined including the any linkage relationships
	  ▪ Questions & Answers (6)

Final Project

You are encouraged to work in a team of 2 for this assignment. If you do, both of you will receive the same grade for this part of the assignment. When you discuss roles, each person should have some individual questions they are responsible for. You may choose to work along, but you will be expected to do the same amount of work as a two-person team.

When you list the questions, have a section on introduction. The introduction should discuss the over arching research topic you are trying to address with your questions (just the tip of the ice burg on this topic, but still), and why studying this with the data is important. I am thinking a short paragraph would suffice. Consider that your target audience is not an expert in your field. So might not know an obvious thing to you. (e.g. Hunger is still a major issue in the US with one in four skipping a meal due to poverty in the US. Don't quote me on this, just remember vaguely something like this). If at all applicable, try to look at multiple years for trends over time.