Verana Health Data Challenge

Thank you for accepting our data challenge! Please answer the questions as well as you can, and send us back an annotated notebook or file in 24 hours from the time you receive this. It should take about 1-3 hours to finish. It's fine if you cannot complete all of the questions. We are hiring for various levels of experience so this would serve as an indicator of how much knowledge you will be coming in with and what you may need to work on. We are interested in your thought processes and what resources you use. Questions were written in order of simple to more complex so it is fine if you cannot finish them all. I suggest reading all the questions first before you start to code. Have fun!

Instructions:

● Use SQL, Python, or PySpark to answer these questions. As much of your code as possible should be executable (Python in particular). If you are not set up to run SQL on your machine, you can write and run the code in Python and write a SQL query that does the same thing for one or two of the questions.

● Write down any assumptions you are making for a question

● You can use any publicly-available resources (ex. Stackoverflow.com, google) to help complete the challenge.

● Note: This dataset is fabricated and not taken from the clinical databases Verana Health leverages.

Database tables (csv format):

patient

procedure

diagnosis

Note: dos = date of service, dx = diagnosis, mod = modifier

Using the included tables, please answer the following questions:

1. Spend some time exploring the data and show or discuss what you find. What are the types of data quality issues or checks should you consider?

2. How many patients have Wet Age-Related Macular Degeneration (wAMD) in the given dataset?

3. How many patients have wAMD in 2019?

4. How many patients have wAMD between 2014-2017, stratified by sex?

5. How would you determine if sex is associated with an increased risk of wAMD?

6. How many women diagnosed with wAMD between 2014-2017 also had an intravitreal injection during that time?

7. What is the most common type of intravitreal injection in women diagnosed with wAMD between 2014-2017?

8. Stratify the type and count of intravitreal injections by eye laterality (right, left, unspecified) in 2014-2017 for patients with wAMD by year.

9. Find the ratio of patient diagnosis dates that have a corresponding procedure date? Does this ratio tell you anything about the data? If so, what might it indicate?

10. Are there any issues with the data? If so, what issues did you notice?

11. How would you define completeness of patient notes and how would you go about validating it?

12. If you were to build a threshold for acceptable data quality, what would you take into consideration and how would you approach it?