INFO 201 DA 2: Pre-Analysis Prep

Analyzing a randomized control trial

In this data assignment, you will reanalyze data that was presented in the paper “Clinical decision support for high-cost imaging: A randomized clinical trial”. The study and analysis described in that paper attempts to estimate how well an information technology system can nudge medical providers (like physicians and nurse practitioners) to make appropriate recommendations about prescribing diagnostic test scans for patients. This type of system is called a clinical decision support or CDS system.

The CDS system is a basic AI system – what is sometimes called an expert decision system that makes recommendations based on a large set of knowledge on a topic provided by experts (rather than systems that base recommendations on machine learning). The way it works is that it is integrated into the computer system that medical providers use when they prescribe or “order” services or treatments for patients. When providers put orders in the system, they must list the reasons for making the orders. In this case, the system is designed to provide “best practice alerts” (BPAs) to providers when they enter orders for diagnostic scans that are considered “inappropriate” given the reasons for ordering them. Diagnostic scans could be things like a PET, CAT, or MRI scan – these scans can help diagnose conditions or identify injuries in some cases, but in many cases they do not provide useful clinical information and are also expensive and sometimes carry extra risk. In cases where some set of symptoms might not warrant having an invasive or expensive diagnostic scan, the CDS system will alert the provider to alternatives that may be more appropriate based on the knowledge base available to the designers of the system. The hope is that implementing such a system will reduce the number of inappropriate scans ordered by medical providers – and in doing so, reduce the financial burden and risk to patients.

For our purposes, we want to estimate the effect of implementing the CDS system on the number of inappropriate scans made by providers. At the same time, we want to check that the CDS system doesn’t also cause unintended negative consequences, such as adding extra burden to providers that result in them serving fewer patients or in quitting their jobs.

Before conducting the analysis itself, you should complete a number of steps!

## Step 1: Familiarize yourself with the source of the data

The data you will examine is an excellent example of using real administrative data to estimate really valuable things in the world. This data comes from a randomized control trial – a fully randomized experiment that is designed to ensure internal validity and pinpoint the causal relationship between a treatment and outcomes. The research problem is at the intersections of AI implementation and information technology, medicine, and economics and its collection were a very complex undertaking. To get a sense of the what the data represents *and* of how difficult it can be to collect data in important domains like healthcare and which require interdisciplinary expertise, you should hear from one of the authors of the original study.

Watch this [IDEA Webinar](https://www.youtube.com/watch?v=Pl1QvLMBdLg) (<https://www.youtube.com/watch?v=Pl1QvLMBdLg>) by Laura Feeney, one of the authors of the study and the Co-Executive Director of the Adbul Latif Jameel Poverty Action Lab, North America. As you watch the video, do the following:

* Try to get a sense of what pressures the research team had in creating the data set in the way that they did – pressures could be related to research goals, ethical considerations, and financial or logistical limits.
* Make a note of every individual person, organization, or team that was necessary to ultimately conduct the study and create the data set. This was a huge undertaking and we are lucky to have access to this data! You will be asked to include a list of people to acknowledge in your assignment! You can make note of the actual names of individuals and also the name of organizations or general teams (different types of analysists or groups of people involved).

Make a list of individuals, organizations, and teams involved in producing the data here! You can reference this alter in making your data report.

## Step 2: Familiarize yourself with the data features

The data that we will look at lists entries for almost every provider in the non-profit Aurora healthcare system (in 2016). For each provider, there are literally hundreds of features recorded. There is data collected before the treatment might have begun: this is called the “quiet period” and is used as a baseline. There is also data collected from the “study period” when the treatment may have started. We will only need to use a small set of these for our analysis. However, you’re going to need to know how to reference these features in our data frame – they will have sometimes uninterpretable names.

To keep track of what each feature means, the data comes with a code book (sometimes called a data dictionary). This code book explains what each feature in the data is intended to represent. In the CDS\_Public\_Use\_File folder, check out CDS\_Documentation and open the file called “cds\_codebook.pdf”.

We will use the data file called “prov\_all\_stats\_sp.dta”. There are 293 features in this data file. Look in the code book to find the names of each of the following features.

* Feature name: ?
  + Whether or not the provider is in the treated group
  + Should tell us 0 if a provider didn’t use the CDS system and 1 otherwise
* Feature name: ?
  + The number targeted scans (in the full study period)
  + These are the “inappropriate” scans where a provider might receive a BPA (best practice alert).
  + We aren’t going to break scans up into subtypes or into the different quarters of the year, just find the feature that includes all scans in the study period (not the quiet period). In general, the study period is the “default” period in the data – and the quiet period data is marked with a q or qp.
* Feature name: ?
  + The number targeted scans (in the full quiet period)
  + These are the “inappropriate” scans where a provider might receive a BPA (best practice alert).
  + We aren’t going to break scans up into subtypes or into the different quarters of the year, just find the feature that includes all scans in the quiet period (not the study period). In general, the study period is the “default” period in the data – and the quiet period data is marked with a q or qp.
* Feature name: ?
  + The number encounters (in the full study period)
  + This is the total number of times that a provider saw patients during the study period.
  + We don’t need to break it up by type of patient or time period within the study period
* Feature name: ?
  + The QP (full quiet period) number encounters
  + This is the total number of times that a provider saw patients during the quiet period.
  + We don’t need to break it up by type of patient or time period within the quiet period
* Feature name: ?
  + Years since provider graduation
  + This should say how many years it has been since a provider received a medical degree (an MD in some cases, but other degree types count for nurse practitioners and other types of providers).
* Feature name: ?
  + Whether the provider had most QP encounters in the ED (emergency department)
  + If most of a provider’s encounters in the quiet period occurred in the emergency department of a hospital, then they may work primarily with critical, urgent, and rapid-pace clinical scenarios
  + We won’t check exactly how much of the time a provider spends in the ED for this analysis, we just want a 1 if the provider is primarily an emergency provider and a 0 otherwise.
* Feature name: ?
  + Whether the provider left health system in the study period
  + Basically: Did the provider quit, get fired, or pass away during the main part of the study
  + We don’t need to know any more detail about timing, just if they did or did not leave the system during the full study period
* Feature name: ?
  + Whether the provider left health system in the quiet period
  + Basically: Did the provider quit, get fired, or pass away during the baseline period before the main part of the study even started
  + We don’t need to know any more detail about timing, just if they did or did not leave the system during the full quiet period

## Step 3: Plan metrics and statistics

### Choice of summary statistic for the success metric

In this analysis, our main outcome of interest is the degree to which the CDS system might reduce the number of “inappropriate” scans (BPA or target scans) that providers order. We’ll look at how the number changes from the quiet period to the study period (after the treatment starts). The idea here is that each inappropriate scan may represent a financial or risk burden to patients. We want to reduce this. Importantly, the goal of this non-profit hospital system is *not* to “sell” as many services as possible. We want to reduce the number of these scans for patients in general or in total in the healthcare system.

Given this goal, we should consider how we formulate our success metric. We will look at a treatment effect. However, we need to decide if an average treatment effect or the median treatment effect makes more sense for our change in scans analysis. The average treatment effect would be the difference in the *mean* change in scans between those who used the CDS and those who didn’t. The median treatment effect would be the difference in *medians*. It is up to you to choose the appropriate option that you will use in the data analysis (and you’ll provide your justification in the data report as well)

Consider:

* Are we interested in only the effect for “typical” providers or do we care about representing “outlier” providers as well?
* Can effects be “pooled” and combined?

For the success metric, which do you think is more appropriate?

* Average treatment effect / Median treatment effect

In about 1-2 sentences, justify why this is the better approach to summarize the data.

* ….

### Choice of summary statistic for the guardrail metric

In addition to checking that the treatment *helps* what we want it to help, we also want to make sure that it doesn’t *hurt* what we don’t want to hurt. In this analysis, you’ll look at two “guardrail” metrics: the number of providers who leave the healthcare system and the change in the number of patient encounters. Here, you should consider the number of patient encounters. If the use of the CDS is really onerous, time-consuming, or annoying, then providers might have less time to actually meet patients. Or alternatively, providers might just avoid seeing patients to avoid having to work with the system. We want to check that this isn’t the case. If we find that providers who use the CSD start seeing substantially fewer patients, then that’s a serious problem! We want to make sure that the number of patient encounters doesn’t go substantially down – it could stay the same, go down just a tiny bit, or go up. In reality, what we really care about isn’t the number of patients that each provider encounters, but the total number of patients that the healthcare system is able to treat.

Given this goal, does it make more sense to consider the average treatment effect of CDS on the change in number of patient encounters or the median treatment effect of CDS on the change in number of patient encounters?

Again, consider:

* Are we interested in only the effect for “typical” providers or do we care about representing “outlier” providers as well?
* Can effects be “pooled” and combined?
* There is no reason why this metric needs to have either the same or a different stat from that used for the success metric.

For the guardrail metric, which do you think is more appropriate?

* Average treatment effect / Median treatment effect

In about 1-2 sentences, justify why this is the better approach to summarize the data.

* ….

## Step 4: Start the analysis!

Go ahead and start working on the analysis! You will create a data report based by completing INFO370\_DA2\_CDS.qmd. The .qmd file is located in the same folder as the document you are reading now. You should make sure to keep the .qmd there, as it references data files.

Some things to note as you fill in the data report: The text is largely written in the past tense. This is because you are preparing a report that is meant to be read by an audience after you have completed the analysis. Additionally, note that this analysis is not fully completed by you, so the author statement includes you and Dr. Miller Rigoli (if you get substantial help from someone, you can list them as well).

## For more information

Read the abstract and introduction (first 5 paragraphs) of the paper that was published describing the original analysis of this study’s data. The paper is available here: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213373>

Note that the analysis conducted is a little different than what we will conduct. The methods used are a little bit more advanced (but not hugely so), and are a bit more time-consuming because the researchers did a lot more work to check additional potential problems with interpretation than we are checking!