../assets/tikzit/tikzit.sty

The effects of continuing enrollment on the completion of clinical trials

Preliminary Draft

William King

February 21, 2025

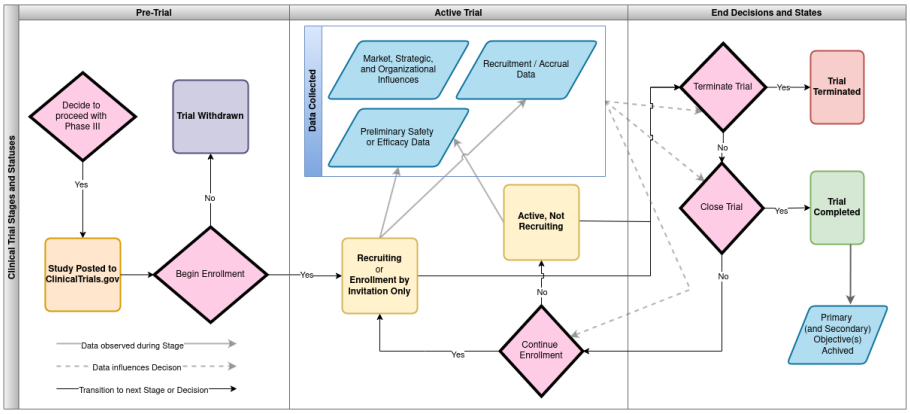
1 Introduction

In 1938, President Franklin D. Roosevelt signed the Food, Drug, and Cosmetic Act, establishing the Food and Drug Administration's (FDA) authority to require pre-market approval of pharmaceuticals [Com14]. This created a regulatory framework where pharmaceutical companies must demonstrate safety and efficacy through clinical trials before bringing drugs to market. The costs of these trials - both in time and money - form a significant barrier to entry in pharmaceutical markets. Understanding what causes clinical trials to fail is therefore crucial to predict the impact of policies, intended or unintended.

Existing research has examined how drugs progress through development pipelines, but we know relatively little about the relative contribution of different challenges to the early termination of clinical trials. When a trial terminates early due to operational challenges rather than safety or efficacy concerns, potentially effective treatments may be delayed or abandoned entirely.

This paper provides the first empirical framework to separate market-driven and safety/efficacy based terminations from one form of operational failure – enrollment challenges – in Phase III clinical trials. Using a novel data-set constructed from administrative data registered on ClinicalTrials.gov, I exploit variation in enrollment timing and market conditions to identify how extending the enrollment period affects trial completion. Specifically, I answer the question: *"How does the probability of trial termination change when the enrollment period is extended?"* This approach differs from previous work that focuses for the most part on the drug development pipeline and progression between clinical trial phases.

To understand how I do this, we'll cover some background information on clinical trials, the current literature, and the administrative data I collected in section 2. Then I'll explain the approach to causal identification and how the data collected matches those results, 3. Then we'll cover the econometric model (section 4) and results (section 5). Finally, we acknowledge deficiencies in the analysis and potential improvements in section 6, then end with my thoughts in the conclusion 7



Diamonds represent decision points while Squares represent states of the clinical trial and Rhombuses represent data obtained by the trial.

Figure 1: Clinical Trial Stages and Progression

2 Clinical Trial Background

To understand why clinical trials succeed or fail requires understanding how they operate and how their progress is documented. The primary source of this operational data is ClinicalTrials.gov, where investigators record key information about their trials' status and progression. To understand how my administrative data captures trial progression, we'll examine how investigators document their trials' states and transitions. Figure 1 is a flowchart of definitions of the different states that a trial can take and the decisions leading to each. It also describes the knowledge obtained by the study operator and how that influences further decisions. The states are standardized and defined by the National Library of Medicine [US 24]. During the prior to a study, the trial investigators will design the trial, choose primary and secondary objectives, and decide on how many participants they need to enroll. Once they have decided on these details, they post the trial to `ClinicalTrials.com` and decide on a date to begin enrolling trial participants. After a trial has enrolled enough participants, the sponsor will move to an "Active, not recruiting" state to inform potential participants that they have recruiting. During this time, the trial operators continue monitoring participants for adverse events and tracking their disease severity and compliance with treatment. Finally, when the investigators have obtained enough data to achieve their primary objective, the clinical trial will be closed and marked as "Completed" in `ClinicalTrials.gov` If the trial is closed before achieving the primary objective, the trial is marked as "Terminated" on `ClinicalTrials.gov`. Trials can be terminated because safety or efficacy evidence suggested it was not worth continuing, enrollment rates were too low to achieve the primary objective within time and budget constraints.

As a trial goes through the different stages of recruitment, the investigators

update the records on ClinicalTrials.gov. Even though there are only a few times that investigators are required to update this information, it tends to be updated somewhat regularly during enrollment as it is a way to communicate with potential enrollees. When a trial is first posted, it includes information such as planned enrollment, planned end dates, the sites at which it is being conducted, the diseases that it is investigating, the drugs or other treatments that will be used, and who is sponsoring the trial. As enrollment is opened and closed and sites are added or removed, investigators will update the status and information to help doctors and potential participants understand whether they should apply.

When a trial ends, it can end in one of three ways. The most desirable outcome is completion, where the trial achieves its primary objective by gathering sufficient data about safety and efficacy. However, trials may also end early either through withdrawal (as mentioned previously) or termination. Termination occurs after enrollment has begun but before achieving the primary objective.

Understanding why trials terminate early is the key goal of this work, but is not straightforward. Terminated trials typically record a description of *a single* reason for the clinical trial termination. This doesn't necessarily list all the reasons contributing to the trial termination and may not exist for a given trial. As an example, if a Principal Investigator leaves for another institution (terminating the trial), this decision may be affected by things such as a safety or efficacy concern, a new competitor on the market, difficulties recruiting participants, or a lack of financial support from the study sponsor. In this way, the stated reason may mask the underlying challenges that led to the termination, leaving us to use another way to infer the relative impact of operational difficulties.

move the following

To better describe termination causes, I suggest classifying them into three broad categories. The first category, Safety or Efficacy concerns, occurs when data suggests the treatment is unsafe or unlikely to achieve its therapeutic goals. While Khmelnitskaya [Khm21] describes these as scientific failures, I contend that they represent successful knowledge gathering - the clinical trial process working as intended to identify ineffective treatments. The second category, Strategic concerns, encompasses business and market-driven decisions such as changes in company priorities or competitive landscape. The final category, Operational concerns, includes practical challenges like insufficient enrollment rates or loss of key personnel. These latter two categories represent true failures of the trial process, as they prevent us from learning whether the treatment would have been safe and effective.

2.1 Literature on Clinical Trials

Clinical trials are a required part of drug development. Not only does the FDA require that a series of clinical trials demonstrate sufficient safety and efficacy of a novel pharmaceutical compound or device, producers of derivative medicines may be required to ensure that their generic small molecule compound – such as ibuprofen or levothyroxine – matches the performance of the originator drug if

delivery or dosage is changed. For large molecule generics (termed biosimilars) such as Adalimumab (Brand name Humira, with biosimilars Abrilada, Amjevita, Cyltezo, Hadlima, Hulio, Hyrimoz, Idacio, Simlandi, Yuflyma, and Yusimry), the biosimilars are required to prove they have similar efficacy and safety to the reference drug.

In the world of drug development, these trials are classified into different phases of development[1]. Pre-clinical studies primarily establish toxicity and potential dosing levels. Phase I trials are the first attempt to evaluate safety and efficacy in humans. Participants typically are healthy individuals, and they measure how the drug affects healthy bodies, potential side effects, and adjust dosing levels. Sample sizes are often less than 100 participants. Phase II trials typically involve a few hundred participants and is where investigators will dial in dosing, research methods, and safety. A Phase III trial is the final trial before approval by the FDA, and is where the investigator must demonstrate safety and efficacy with a large number of participants, usually on the order of hundreds or thousands. Occasionally, a trial will be a multi-phase trial, covering aspects of either Phases I and II or Phases II and III. After a successful Phase III trial, the sponsor will decide whether or not to submit an application for approval from the FDA. Before filing this application, the developer must have completed "two large, controlled clinical trials." Phase IV trials are used after the drug has received marketing approval to validate safety and efficacy in the general populace. Throughout this whole process, the FDA is available to assist in decision-making regarding topics such as study design, document review, and whether they should terminate the trial. The FDA also reserves the right to place a hold on the clinical trial for safety or other operational concerns, although this is rare. [Com28].

In the economics literature, most of the focus has been on describing how drug candidates transition between different phases and their probability of final approval. Abrantes-Metz, Adams, and Metz [AAM04] described the relationship between various drug characteristics and how the drug progressed through clinical trials. They found that as Phase I and II trials last longer, the rate of failure increases. In contrast, Phase 3 trials generally have a higher rate of success than failure after 91 months. This may be due to the fact that the purpose of Phases I and II are different from the purpose of Phase III.

Continuing on this theme, DiMasi et al. [DiM+10] examine the completion rate of clinical drug development and find that for the 50 largest drug producers, approximately 19% of their drugs under development between 1993 and 2004 successfully moved from Phase I to receiving an New Drug Application (NDA) or Biologics License Application (BLA). They note a couple of changes in how drugs are developed over the years they study, most notably that drugs began to fail earlier in their development cycle in the latter half of the time they studied. They note that this may reduce the cost of new drugs by eliminating late and costly failures in the development pipeline.

Earlier work by DiMasi [DiM02] used data on 68 investigational drugs from

[1] [22] provide an overview of this process while [Com28] describes the process in detail.

10 firms to simulate how reducing time in development reduces the costs of developing drugs. He estimates that reducing Phase III of clinical trials by one year would reduce total costs by about 8.9% and that moving 5% of clinical trial failures from phase III to Phase II would reduce out of pocket costs by 5.6%.

A key contribution to this drug development literature is the work by Khmelnitskaya [Khm21] who created a causal identification strategy to disentangle strategic exits from exits due to clinical failures in the drug development pipeline. She found that overall 8.4% of all pipeline exits are due to strategic terminations and that the rate of new drug production would be about 23% higher if those strategic terminatations were eliminated.

The work that is closest to mine is the work by Hwang et al. [Hwa+16] who investigated causes for which late stage (Phase III) clinical trials fail – with a focus on trials in the USA, Europe, Japan, Canada, and Australia. They identified 640 novel therapies and then studied each therapy's development history, as outlined in commercial datasets. They found that for late stage trials that did not go on to receive approval, 57% failed on efficacy grounds, 17% failed on safety grounds, and 22% failed on commercial or other grounds.

Unfortunately the work of both Hwang et al. [Hwa+16] and Khmelnitskaya [Khm21] ignore a potentially large cause of failures: operational challenges, i.e. when issues running or funding the trial cause it to fail before achieving its primary objective. In a personal review of 199 randomly selected clinical trials which terminated before achieving their primary objective, I found that 14.5% cited safety or efficacy concerns, 9.1% cited funding problems (an operational concern), and 31% cited enrollment issues (a separate operational concern)[2].

2.2 Introduction to ClinicalTrials.Gov

Since Sep 27th, 2007 those who conduct clinical trials of FDA controlled drugs or devices on human subjects must register their trial at `ClinicalTrials.gov` ([22]). This involves submitting information on the expected enrollment and duration of trials, drugs or devices that will be used, treatment protocols and study arms, as well as contact information the trial sponsor and treatment sites.

When starting a new trial, the required information must be submitted "…not later than 21 calendar days after enrolling the first human subject…". After the initial submission, the data is briefly reviewed for quality and then the trial record is published and the trial is assigned a National Clinical Trial (NCT) identifier. ([22]).

Each trial's record is updated periodically, including a final update that must occur within a year of completing the primary objective, although exceptions are available for trials related to drug approvals or for trials with secondary objectives that require further observation[3] ([22]). Other than the requirements for the first and last submissions, all other updates occur at the discretion of the trial sponsor. Because the ClinicalTrials.gov website serves as a central point

[2]Note that these figures differ from Hwang et al. [Hwa+16] because I sampled from all stages of trials, not just Phase III trials focused on drug development.

[3]This rule came into effect in 2017

of information on which trials are active or recruiting for a given condition or drug, most trials are updated multiple times during their progression.

There are two primary ways to access data about clinical trials. The first is to search individual trials on ClinicalTrials.gov with a web browser. This web portal shows the current information about the trial and provides access to snapshots of previously submitted information. Together, these features fulfill most of the needs of those seeking to join a clinical trial. For this project I've been able to scrape these historical records to establish snapshots of the records provided. The second way to access the data is through a normalized database setup by the Clinical Trials Transformation Initiative called AACT. The AACT database is available as a PostgreSQL database dump or set of flat-files. These dumps match a near-current version of the ClinicalTrials.gov database. This format is amenable to large scale analysis, but does not contain information about the past state of trials. I combined these two sources, using the AACT data-set to select trials of interest and then scraping `ClinicalTrials.gov` to get a timeline of each trial. The result is a series of snapshots, each documenting a specific set of recorded changes in a trial. It is these snapshots that provide the opportunity to estimate the data generating process corresponding to the clinical trials for which I have data.

3 Causal Story and Data

As I am trying to separate strategic concerns (the effect of a marginal treatment methodology) and an operational concern (the effect of a delay in closing enrollment), we need to look at what confounds these effects and how we might measure them. To start, we'll look at the data generating model, the values of interest, and both the observed and unobserved confounding variables. We'll also discuss how the data collected fits the data generating process.

The primary effects one might expect to see are that

1. Adding more drugs to the market will make it harder to finish a trial as it is more likely to be terminated due to concerns about profitability.
2. Adding more drugs to the market will make it harder to recruit, slowing enrollment.
3. Enrollment challenges (i.e. delays) increase the likelihood that a trial will terminate.

Unfortunately, these causal effects are confounded in many different ways. Figure 2 contains a description of the causal model.

3.1 Causal Identification

Because running experiments on companies running clinical trials is not going to happen anytime soon, causal identification depends on using a structural causal model. Because the data generating process for the clinical trials records

is rather straightforward, this is an ideal place to use Pearl [Pea09] Do-Calculus. This process involves describing the data generating process in the form of a directed acyclic graph, where the nodes represent different variables within the causal model and the directed edges (arrows) represent assumptions about which variables influence the other variables. There are a few algorithms that then tell the researcher which of the relationships will be confounded, which ones can be statistically estimated, and provides some hypotheses that can be tested to ensure the model is reasonably correct.

In **??** I diagram the directed acyclic graph that describes my proposed data generating process, It revolves around the decisions made by the study sponsor, who must decide whether to let a trial run to completion or terminate the trial early. While receiving updates regarding the status of the trial, they ask questions such as:

- Do I need to terminate the trial due to safety incidents?
- Does it appear that the drug is effective enough to achieve our goals, justifying continuing the trial?
- Are we recruiting enough participants to achieve the statistical results we need in the budget we have?
- Does the current market conditions and expectations about returns on investment justify the expenditures we are making?

When appropriate issues arise, the study sponsor terminates the trial, otherwise it continues to completion.

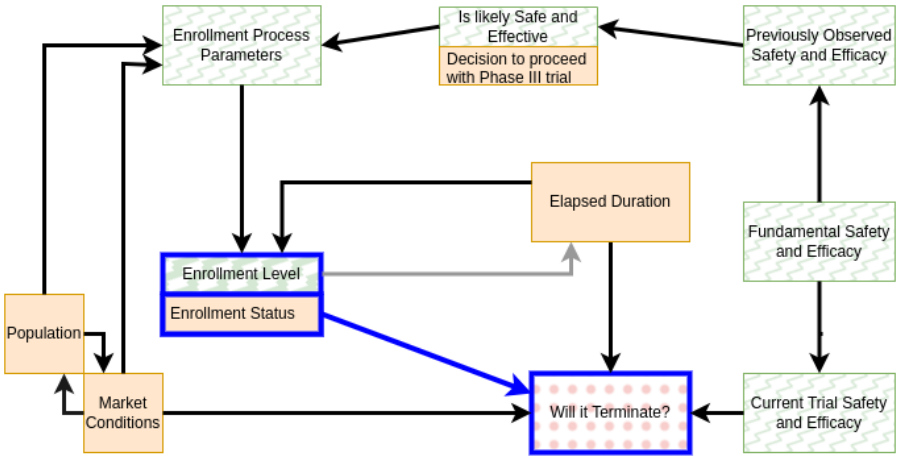


Figure 2: Graphical Causal Model

A quick summary of the nodes of the DAG, which nodes are captured in the data, the hypothesized relationships in the model, and the proposed confounding pathways.

- Items of Interest (Blue boxes and Arrow)
    1. `Enrollment Level (Enrollment Status)`: While occasionally a trial will keep the enrollment numbers up to date, the only regular information on enrollment received is the enrollment status, i.e. whether they have finished recruiting or not.
    2. `Will it Terminate?`: This represents whether the trial was terminated or if it completed successfully.
    3. The effect of `Enrollment Status` on `Will it Terminate?`: How does changing the enrollment status affect the probability of termination.
- Observed values (Solid orange boxes)
    1. `Condition` (Not drawn in DAG because it impacts everything): The underlying condition, classified by ICD-10 group. This impacts every other aspect of the model and is pulled from the AACT data-set.
    2. `Market Measures`: Various measures of the number of alternate drugs on the market. These are either the number of other drugs with the same active ingredient as the trial (both generic and originators), and those considered alternatives in various formularies published by the United States Pharmacopoeia.
    3. `Population (market size)`: Multiple measures of the impact the disease. These are measured by the DALY cost of the disease, and is separated by the impact on countries with High, High-Medium, Medium, Medium-Low, and Low Socio-Demographic Index (SDI) scores. This data comes from the Institute for Health Metrics' Global Burden of Disease study [Vos+20].
    4. `Elapsed Duration`: A normalized measure of the time elapsed in the trial. Comes from the original estimate of the trial's primary completion date and the registered start date. I take the difference in days between these, and get the percentage of that time that has elapsed. This calculation is based on data from the snapshots and the AACT final results.
    5. `Decision to Proceed with Phase III`: If the compound development has progressed to Phase III. This is included in the analysis by only including Phase III trials registered in the AACT data-set.
- Unobserved (Green Boxes with squiggle hatch marks)
    1. `Fundamental Efficacy and Safety`: The underlying safety of the compound. Cannot be observed, only estimated through scientific study.

2. `Previously observed Efficacy and Safety`: The information gathered in previous studies. This is not available in my data-set because I don't have links to prior studies.
3. `Currently observed Efficiency and Safety`: The information gathered during this study. This is only partially available, and so is treated as unavailable. After a study is over, the investigators are often publish information about adverse events, but only those that meet a certain threshold. As this information doesn't appear to be provided to participants, we don't consider it.

- Jointly determined variables

1. `Enrollment Level (Enrollment Status)` $\leftrightarrow$ `Elapsed Duration`: Because I only observe enrollment status and have not good estimate of the enrollment process, there is a potential for confounding between the elapsed duration of a trial and the enrollment status. The proposed mechanisms are through the partially observed levels of enrollment. First, as a trial progresses, the enrollment levels should grow until it matches the planned enrollment and the trial ends. Thus under good circumstances, elapsed duration drives enrollment levels. Under bad circumstances though, low enrollment levels may cause the duration to extend, as study sponsors spend more resources to complete the trial successfully. This is an issue because the only complete measure of enrollment that we currently have is the enrollment status, and thus I cannot control for this effect.
2. `Market Conditions` $\leftrightarrow$ `Population`: There exists an endogenous dynamic between between the treatments available for a disease and the market size/population with that disease. Cerda [Cer07] proposes two mechanisms that link the drugs on the market and market size. The first is that a larger population increases the potential profitability, trying to get more treatments allowed. The inverse is that for many chronic diseases with high mortality rates, more drugs cause better survivability, increasing the size of those markets.

- Confounding Pathways

1. `Condition` (Not drawn in figure 2): Interacts with everything.
2. Backdoor Pathway between `Will Terminate?` and `Enrollment Status` through `Fundamental Safety and Efficacy`. The concern is that since previously learned information and current information are driven by the same underlying physical reality, the enrollment process and termination decisions may be correlated. Controlling for the decision to proceed with the trial is the best adjustment available to block this confounding pathway. Below I describe the exact pathways.

(a) `Will Terminate?` ← `Currently Observed Efficacy and Safety` ← `Fundamental Efficacy and Safety` → `Previously Observed Efficacy and Safety` → `Is likely safe and effective (Decision to proceed with Phase III trial)` → `Enrollment Process Parameters` → `Enrollment Levels (Enrollment Status)`

3. Backdoor Pathways through `Population` and `Market Conditions` The concern with this pathway is that the rate of enrollment, and thus the enrollment status, is affected by the Population with the disease and the market condition.
   (a) `Will Terminate?` ← `Market Conditions` → `Enrollment Process Parameters` → `Enrollment Levels (Enrollment Status)`
   (b) `Will Terminate?` ← `Market Conditions` ↔ `Population` → `Enrollment Process Parameters` → `Enrollment Levels (Enrollment Status)`
4. Backdoor Pathway through `Elapsed Duration`.
   (a) `Will Terminate?` ← `Elapsed Duration` ↔ `Enrollment Levels (Enrollment Status)`

In the sections below, I examine each source of data, their key features, how they match with the variables in the Structural Model DAG, and describe applicable terminology (section 3.2). I then discuss how these sources were tied together (section 3.2.4) and describe the specific data used in the analysis (section 3.3).

3.2 Data Sources

3.2.1 Clinical Trials Data

Since Sep 27th, 2007 those who conduct clinical trials of FDA controlled drugs or devices on human subjects must register their trial at `ClinicalTrials.gov` [US b]. This involves submitting information on the expected enrollment and duration of trials, drugs or devices that will be used, treatment protocols and study arms, as well as contact information the trial sponsor and treatment sites.

When starting a new trial, the required information must be submitted "…not later than 21 calendar days after enrolling the first human subject…". After the initial submission, the data is briefly reviewed for quality and then the trial record is published and the trial is assigned a National Clinical Trial (NCT) identifier. [US b].

Each trial's record is updated periodically, including a final update that must occur within a year of completing the primary objective, although exceptions are available for trials related to drug approvals or for trials with secondary objectives that require further observation[4] [US b]. Other than the requirements for the the first and last submissions, all other updates occur at the discretion of the trial sponsor. Because the ClinicalTrials.gov website serves as a central

[4]This rule came into effect in 2017

point of information on which trials are active or recruiting for a given condition or drug, most trials are updated multiple times during their progression.

There are two primary ways to access data about clinical trials. The first is to search individual trials on ClinicalTrials.gov with a web browser. This web portal shows the current information about the trial and provides access to snapshots of previously submitted information. Together, these features fulfill most of the needs of those seeking to join a clinical trial. The second way to access the data is through a normalized database setup by the Clinical Trials Transformation Initiative (CTTI [Cli22] called the Aggregate Analysis of ClinicalTrials.gov (AACT). The AACT database is available as a PostgreSQL database dump or set of pipe ("|") delimited files and matches the current version of the ClinicalTrials.gov database. This format is amenable to large scale analysis, but does not contain information about the past state of trials.

I created a set of python scripts to incorporate the historical data on clinical trials available through the web portal and merge it into a local copy of the standard AACT database. This novel data-set can be used to easily track changes as trials progresses.

In this combined data-set of current and historical trial records, there are a few areas of particular interest.

- NCT: As a unique identifier of a trial, it is used throughout to ensure data is linked to the appropriate trial.
- Enrollment: This takes on two forms. At the beginning of a trial this is presented as "Anticipated" enrollment, while near or at the end of the trial it is reported as "Actual" enrollment.
- Overall Status: Each trial must be in one of a list of states. While a trial is running, it can be in any of the following states.
  - Not yet recruiting
  - Recruiting
  - Enrolling by Invitation
  - Active, not recruiting
  - Suspended

  When a trial has ended it is in one of three states:
  - Withdrawn: Trial has ended before any enrollment began. I filtered all of these out as they do not apply to our work.
  - Terminated: Trial has ended prematurely.
  - Completed: Trial has ended after observing what they hoped to observe.
- Start Date: The date that the first measurement was taken or that the first site was authorized to take measurements.

- Primary Completion Date: The date the last measurement for the primary objective was taken. Prior to the actual primary completion date, this is an anticipated value.
- Conditions: The conditions of interest in the trial.
- Interventions: The drug(s) used in treatment.

3.2.2 Drug Compounds and Structured Product Labels (SPLs)

When a drug is licensed for sale in the U.S., it is not just the active ingredients that are licensed, but also the dosage and route of administration. Each of these combined compound/dosage/route pairs are assigned a unique National Drug Code (NDC). The list of approved NDCs are released regularly in the FDA's Orangebook (small-molecule drugs) and Purplebook (Biologicals) publications. These two publications also contain information regarding which drugs are generics or biosimilars.

Before a drug or drug compound is sold on the market, the FDA requires the seller to submit a standardized label and associated information called a Structured Product Label (SPL). These SPLs include information about dosage, ingredients, warnings, and the format of the printed labels. Each NDC code can have multiple SPLs associated with it because each drug compound may be packaged in multiple ways, e.g. boxes with different numbers of blister packs, etc. These SPLs are made available for download so that they can be integrated into patient health systems to improve patient safety [US a].

The FDA also published additional data in the NDC SPL Data Elements (NSDE) file. This file contains some of the data from the SPL files, as well as the dates when each product was approved for sale and when it was removed from the market. This summary of SPLs is what I used to find which drugs were approved to be on the market at a given date.

3.2.3 Global Burdens of Disease (2019)

The University of Washington's Institute for Health Metrics and Evaluation published a data-set called the Global Burdens of Disease Study 2019 (GBD 2019). This data-set provides estimates of worldwide incidence of various diseases and classes of diseases. The available measures of incidence include Deaths, Disability Adjusted Life Years (DALYs), Years of Life Lost (YLL), and Years Lived with Disability (YLD) and come with both an estimate and 90% confidence interval bounds. Estimates are available for national, multinational, and global populations [Vos+20].

These classes of disease are organized in a hierarchy, with each subsuming category having its own estimates of disease incidence. One understandable deficiency in this data-set is that it doesn't account for all diseases tracked in other data-sets, but focuses on those that are most important from a public health perspective.

3.2.4 Medical and Pharmacological Terminologies

In order to link these disparate data sources I used multiple standardized terminologies. In each section below I briefly describe each terminology, its contents, and uses.

3.2.4.1 Medical Subject Headings (MeSH) Thesaurus

The Medical Subject Headings (MeSH) Thesaurus is produced and maintained by the National Library of Medicine. It is used to index subjects in various NLM publications including PubMed [US c]. The AACT database contains a table that links clinical trials' clinical conditions and drug names to terms in the MeSH thesaurus. As this contains a standardized nomenclature, it simplified much of the linking between clinical trials and other data sources.

3.2.4.2 RxNorm

According to [US e]

> What is RxNorm?
> RxNorm is two things: a normalized naming system for generic and branded drugs; and a tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems....

Both of these functions are crucial to the analysis. The normalized naming system allowed me to convert a diverse set of names as recorded for each clinical trial into standardized identifiers. These standardized identifiers are known as RxCUIs, and they are used in RxNorm to identify not only individual drug components, but also brand names, licensed drug/dosage pairs, and packages. The links to other drug terminologies included links to SPL identifiers, which permitted me to link each trial to drugs on the market at and point in time.

The RxNorm data is provided in multiple formats. The one I chose to use was a MariaDB database that backs a service called RxNav provided by the National Library of Medicine (NLM). The NLM provides scripts to set up and host the backing databases on your own servers [US d]. After setting up the local server, I wrote a python program to export the data from the RxNorm database and import it into the AACT Database. This was required because the former uses a MariaDB database server and the latter uses a Postgres database server.

With the data now available alongside the AACT database, I could link trials to various key drug concepts, including normalized drug ingredient names, NDCs incorporating those ingredients, and the brand names associated with the NDCs.

3.2.4.3 International Classification of Diseases 10th revision (ICD-10)

The International Classification of Diseases 10th revision (ICD-10) is a worldwide standard for categorizing human disease maintained by the World Health Organization. Although the WHO version's last major update was in 2019 and it was officially superseded in 2022 by the 11th revision [Worb] . the 10th revision is still in use in the United States as the Centers for Medicare and Medicaid Services (CMS) continues to publish updated versions called ICD-10-CM (Clinical Management) [Cen22b] and ICD-10-PCS (Procedure Coding System) [Cen22a] for use in medical billing.

ICD-10 codes are organized in a hierarchy. There are 22 highest level categories, representing general categories such as cancers, mental illness, and infectious diseases. The second layer of the hierarchy consists of about 225 groupings.

As I needed a combined list of ICD-10 codes, I first obtained the 2019 version of the ICD-10-CM codes from the CMS (the most recent version corresponding to the GBD matching file) With the arrival of the ICD-11 system, it was difficult to find an official source from which to download the WHO versions of ICD-10 codes. Eventually I resorted to copying them from the navigation bar of the [Wora]. After getting both sources into the same format, I combined them and removed duplicate codes, preferring to keep the descriptions from the WHO version. This was done using standard UNIX scripting commands. I then imported the data into the Postgres Database alongside the AACT data.

3.3 Data Integration

Below is more information about how the data was used in the analysis.

For clinical trials, I captured each update that occurred after the start date and prior to the primary completion date of the trial. For clarity I will refer to these as a snapshot of the trial.

For each snapshot I recorded the enrollment (actual or anticipated), the date the it was submitted, the planned primary completion date, and the trial's overall status at the time. I also extracted the anticipated enrollment closest to the actual start date of the trial, which I will call the planned enrollment under the assumption that the sponsor is recording their current plan for enrollment.

I then calculated a normalized measure of how far along the trial was in it's planned duration; in other word, a measure of elapsed duration. This was calculated for each snapshot as:

$$\text{Elapsed Duration} = \frac{\text{Snapshot Date} - \text{Start Date}}{\text{Primary Completion Date (anticipated)} - \text{Start Date}} \tag{1}$$

Note that this has a range of $[0, \infty)$ although for practical matters it is only about $[0, 3]$. I also included the current status by encoding it to dummy parameters.

As an initial measure of market conditions I have gathered the number of brands that are producing drugs containing the compound(s) of interest in the trial. This was done by extracting the RxCUIs that represented the drugs of interest, then linking those to the RxCUIs that are brands containing those ingredients. As a secondary measure of market conditions, I linked clinical trials to the USP Drug Classification list.

Once I had linked the drugs used in a trial to the applicable USP DC category and class, I could find the number of alternative brands in that class. This matching was performed by hand, using a custom web interface I wrote. In order to link clinical trials to standardized ICD-10 conditions and thus to the Global Burdens of Disease Data, I wrote a python script to search the UMLS system for ICD-10 codes that matched the MeSH descriptions for each trial. This search resulted in generally three categories of search results:

1. The results contained a few entries, one of which was obviously correct.
2. The results contained a large number of entries, a few of which were correct.
3. The results did not contain any matches.

After manually matching each trial to an ICD-10 code, each trial is easily linked to either one of the 22 highest level categories or the 225 or so 2nd level categories in the ICD-10 hierarchy. Linking to one of the disease categories in the GBD hierarchy is similarly easy. To get the best estimate of the size of the population associated with a disease, each trial is linked to the most specific disease category applicable. As not every ICD-10 code is linked to a condition in the GBD, those without any applicable conditions were dropped from the data-set.

4 Econometric Model

The goal is to take each snapshot and predict the probability of termination. To this end, the model I use is a hierarchical logistic regression model where the hierarchies correspond to the 22 top-level ICD-10 disease categories.

First, some notation:

- i: indexes trials
- n: indexes trial snapshots.
- y_i: whether each trial terminated (true, 1) or completed (false, 0).
- d_i: indexes the ICD-10 disease category of the trial.
- $x_{i,n}$: represents the independent variables associated with the snapshot.

The specification of the model to measure the direct effect of enrollment status is:

$$y_i \sim \text{Bernoulli}(p_{i,n}) \tag{2}$$

$$p_{i,n} = \text{logit}(x_{i,n}\vec{\beta}(d_i)) \tag{3}$$

Where beta is indexed by $d \in \{1, 2, \ldots, 21, 22\}$ for each general ICD-10 category. The βs are distributed

$$\beta(d_i) \sim \text{Normal}(\mu_i, \sigma_i I) \tag{4}$$

With hyper-priors

$$\mu_k \sim \text{Normal}(0, 0.05) \tag{5}$$

$$\sigma_k \sim \text{LogNormal}(-2.1, 0.2) \tag{6}$$

The independent variables include:

$$\begin{aligned}
x_{i,n}\beta(d_i) =& \beta_1(d_i) \times \text{Elapsed Duration} && \text{(7a)}\\
&+ \beta_2(d_i) \times \text{arcsinh}\,(\#\text{ USP-DC alternate compounds}) && \text{(7b)}\\
&+ \beta_3(d_i) \times \text{arcsinh}\,(\#\text{ Brands with same compound}) && \text{(7c)}\\
&+ \beta_4(d_i) \times \text{arcsinh}\,(\#\text{ DALYs in High SDI Countries}) && \text{(7d)}\\
&+ \beta_5(d_i) \times \text{arcsinh}\,(\#\text{ DALYs in High-Medium SDI Countries}) && \text{(7e)}\\
&+ \beta_6(d_i) \times \text{arcsinh}\,(\#\text{ DALYs in Medium SDI Countries}) && \text{(7f)}\\
&+ \beta_7(d_i) \times \text{arcsinh}\,(\#\text{ DALYs in Low-Medium SDI Countries}) && \text{(7g)}\\
&+ \beta_8(d_i) \times \text{arcsinh}\,(\#\text{ DALYs in Low SDI Countries}) && \text{(7h)}\\
&+ \beta_9(d_i) \times I_{\text{Not yet Recruiting}}(\text{Trial Status}) && \text{(7i)}\\
&+ \beta_{10}(d_i) \times I_{\text{Enrolling by Invitation Only}}(\text{Trial Status}) && \text{(7j)}\\
&+ \beta_{11}(d_i) \times I_{\text{Recruiting}}(\text{Trial Status}) && \text{(7k)}\\
&+ \beta_{12}(d_i) \times I_{\text{Active, not recruiting}}(\text{Trial Status}) && \text{(7l)}
\end{aligned}$$

Note that the last four are the enrollment status fixed effects. I used the arcsinh because it is similar to a log transform but differentiably handles counts of zero[5]. Some of the other variables are implicitly controlled for as they are used to select the trials of interest. These include:

- The trial is Phase 3.
- The trial has a Data Monitoring Committee[6].
- The compounds are FDA regulated drugs.
- The trial was never suspended[7]

[5]For those unfamiliar, $\text{arcsinh}(x) = \ln(x + \sqrt{x^2 + 1})$.

[6]Trials that are involved in drug development typically have a DMC.

[7]This was because I wasn't sure how to handle it in the model when I started scraping the data.

4.1 Interpretation

The specific measure of interest is how much the probability of terminating a trial changes when there is an delay in closing enrollment, exogenous to timing. Due to the Bayesian nature of the analysis, there is a useful tool to analyze the effect of the delay, called a distribution of differences plot. In the standard reduced form causal inference, the treatment effect of interest for outcome Z is measured as

$$E(Z(\text{Treatment}) - Z(\text{Control})) = E(Z(\text{Treatment})) - E(Z(\text{Control})) \quad (8)$$

Because $Z(\text{Treatment})$ and $Z(\text{Control})$ are random variables, $Z(\text{Treatment}) - Z(\text{Control}) = \delta_Z$, is also a random variable. In the Bayesian framework, this parameter has a distribution, and so we can calculate the distribution of differences in the probability of termination between treatment and control $p_{i,n}(T) - p_{i,n}(C) = \delta_{p_{i,n}}$. This not only gives the average treatment effect ($E[\delta_{p_{i,n}}]$) but all other summaries of the treatment effect can be calculated from this distribution.

I calculate this posterior distribution of $\delta_{p_{i,n}}$ by estimating the posterior distributions of the βs and then simulating $\delta_{p_{i,n}}$. This involves taking a draw from the βs distribution, calculating $p_{i,n}(C)$ for the underlying trials at the snapshot when the trial closed enrollment (changed status to "Active, not recruiting") and then calculating $p_{i,n}(T)$ under the counterfactual where enrollment had not yet closed (accomplished by changing the status to "Recruiting"). This is, in effect, an instantaneous and exogenous extension of the enrollment period. The difference $\delta_{p_{i,n}}$ is then calculated for each trial and parameter sample. After repeating this for all the posterior samples and all trials at their point of close, we have an estimate for the posterior distribution of differences between treatment and control for selected trials. From this distribution I then plot and calculate summaries to assist in interpreting the results.

5 Results

In this section I describe the model fitting, the posteriors of the parameters of interest, and interpret the results.

5.1 Data Summaries and Estimation Procedure

Overall, I successfully processed 168 trials, with 1,347 snapshots between them. Figure 3 shows the histogram of snapshots per trial. Most trials lasted less than 1,500 days, as can be seen in 4. Although there are a large number of snapshots that will be used to fit the model, the number of trials – the unit of observation – are quite low. This is compounded by the fact that these are spread over multiple ICD-10 categories as can be seen in figure 5. The hierarchical approach helps because it pools information between categories.

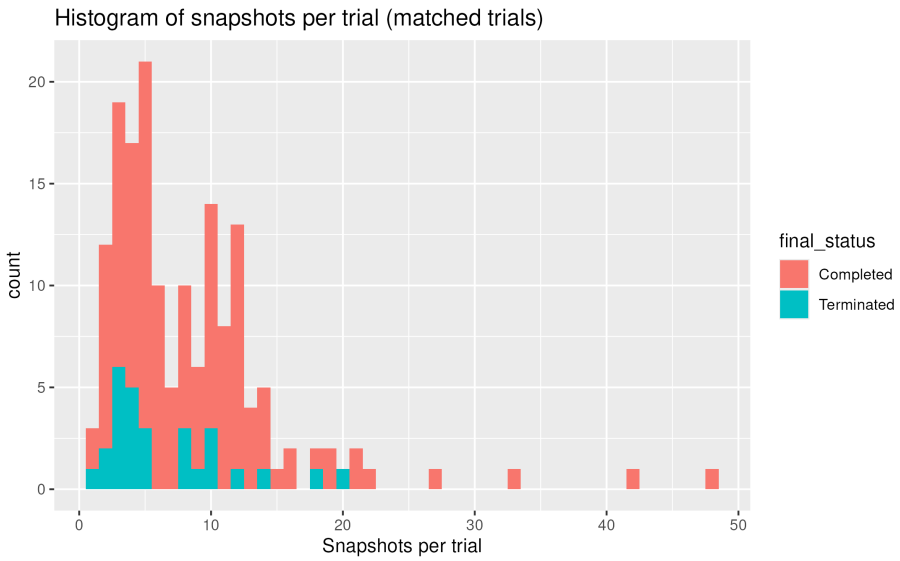


Figure 3: Histogram of the count of Snapshots

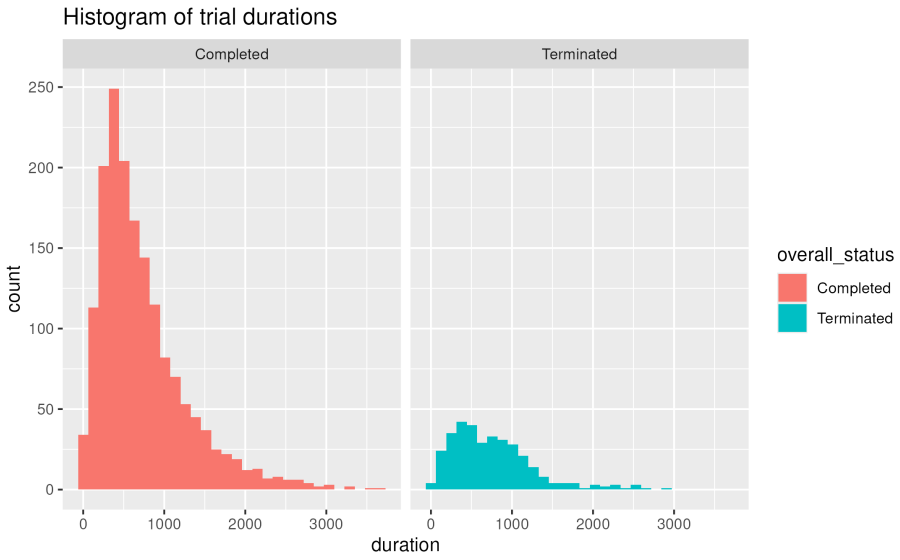


Figure 4: Histograms of Trial Durations

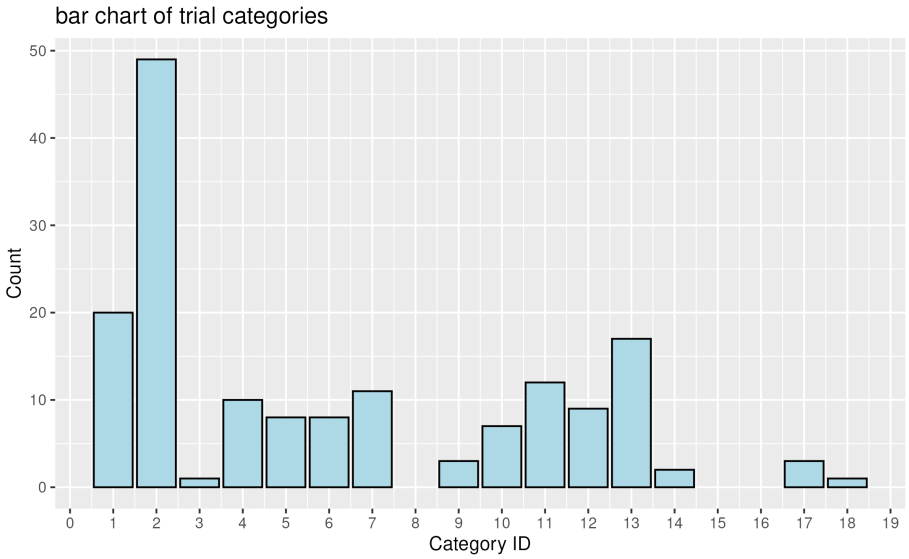


Figure 5: Bar chart of trials by ICD-10 categories

We can use a scatterplot to get a rough idea of the observed relationship between the number of snapshots and the duration of trials. We can see this in Figure 6, where the correlation (measured at 0.34) is apparent.

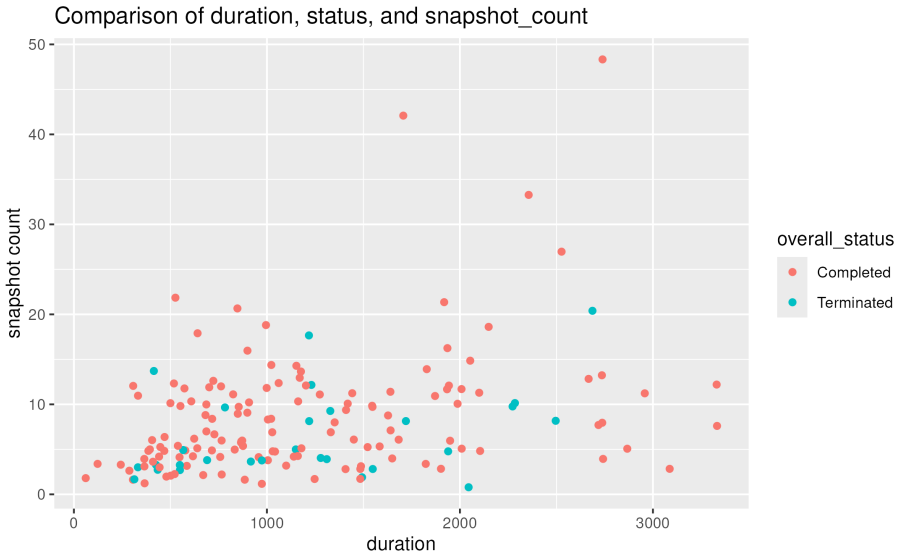
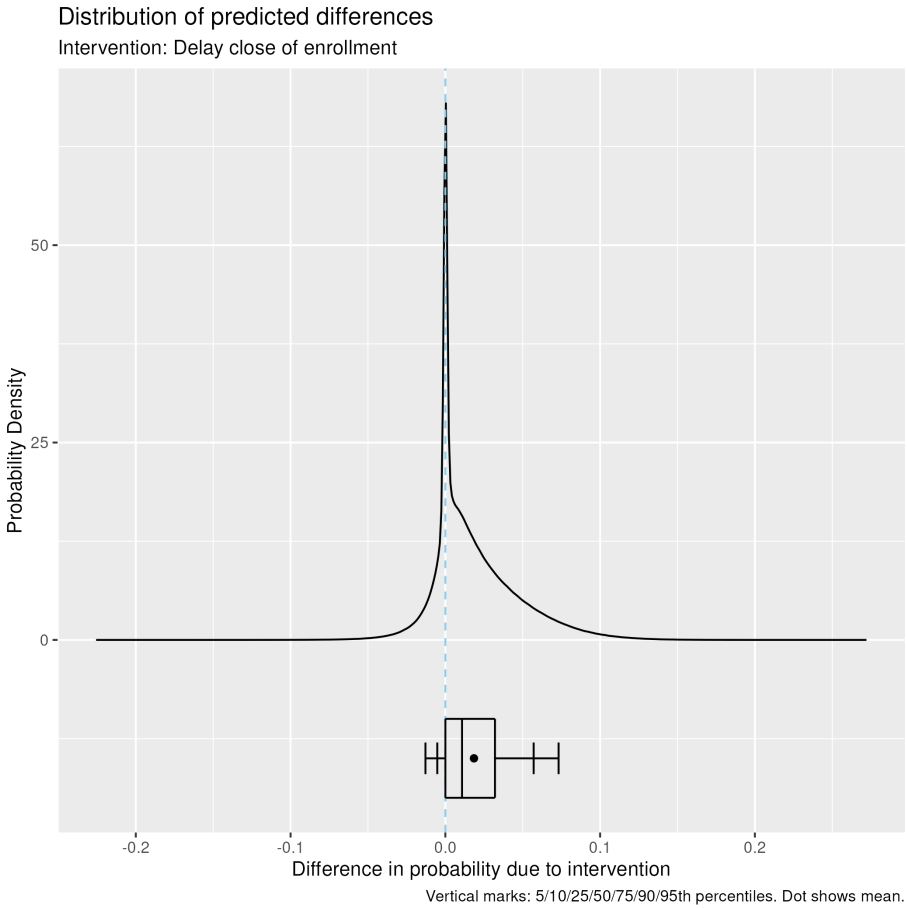


Figure 6: Scatterplot comparing the Count of Snapshots and Trial Duration

I fit the econometric model using mc-stan [Sta22] through the rstan [Sta23] interface using 8 chains with 4,000 warm-up iterations and 8,000 sampling iterations each. No convergence warnings were issued.

5.2 Primary Results

The primary, causally-identified value we can estimate is the change in the probability of termination caused by (counterfactually) keeping enrollment open instead of closing enrollment when observed. In figure 7 below, we see this impact of keeping enrollment open.



Values near 1 indicate a near perfect increase in the probability of termination. Values near 0 indicate little change in probability, while values near -1, represent a decrease in the probability of termination. The scale is in probability points, thus a value near 1 is a change from unlikely to terminate under control, to highly likely to terminate.

Figure 7: Histogram of the Distribution of Predicted Differences

Table 1: Boxplot Summary Statistics: percentage point due to intervention

5th	10th	25th	median	75th	90th	95th	mean
-2.1	-0.8	0.0	1.2	4.2	8.2	11.0	2.5

The key figures from the boxplot in figure 7 are summarized in table 1 There are a few interesting things to point out here. First, approximately 75% of the probability mass is equal to or above zero, suggesting that in most cases a trial will experience some harm from a delay in closing enrollment. Second, the average treatment effect is to increase the probability of termination by about

2.5 percentage points. The full 5% percentile table can be found in table 2 in appendix A

Figure 8 shows how the different disease categories tend to have a similar distribution of differences.

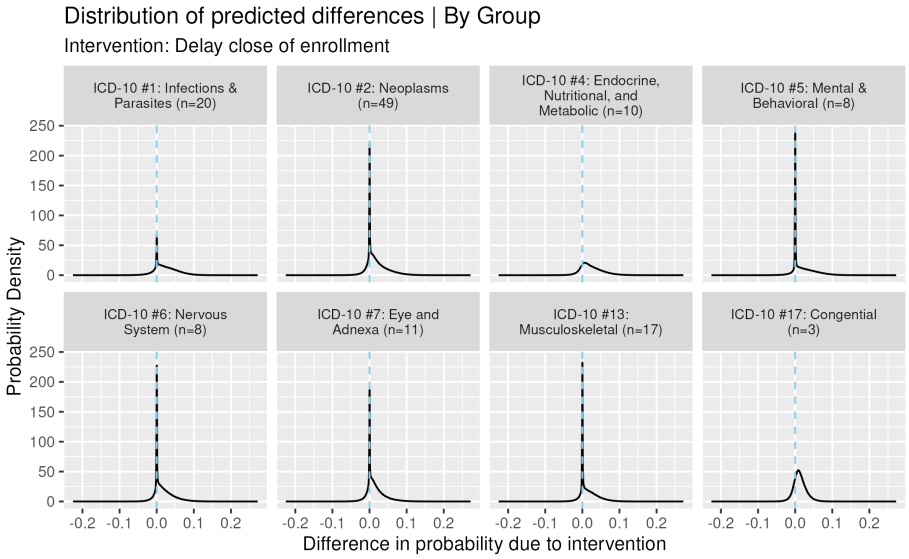


Figure 8: Distribution of Predicted differences by Disease Group

Although these distributions quite similar to each other, we can observe the differences between the fixed effects of the "Active, not recruiting" and "Recruiting" statuses in figure 9. For categories with relatively high numbers of observations (e.g. Neoplasms $n = 49$ or Infections and Parasites $n = 20$) we see a lot of movement, but for categories without any observations (e.g. Special Purposes, Contact with Healthcare etc.) we observe the effects of the implicit pooling in the model, where the probability of termination increases. Table 3 in appendix A contains the summary statistics for each of these distributions

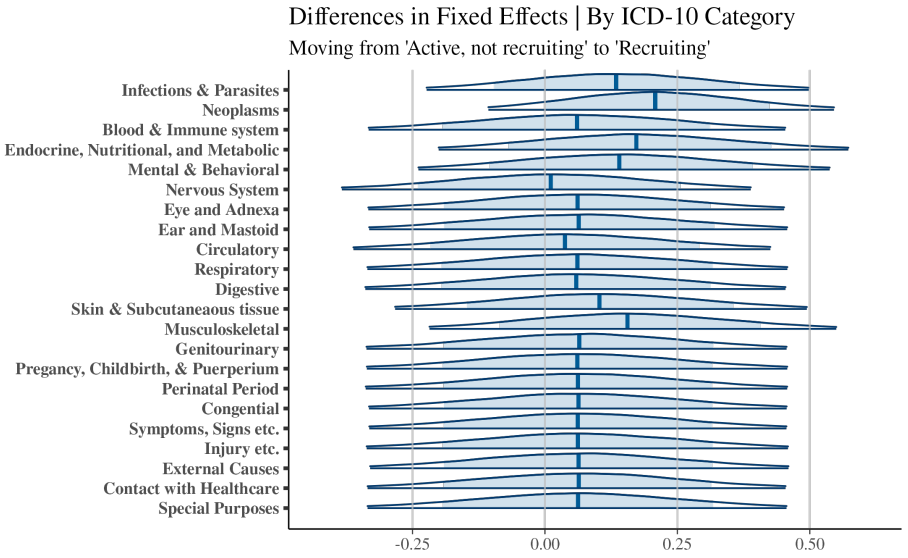


Figure 9: Relative fixed effect of moving from "Active, not recruiting" to"Recruiting" status, by ICD-10 Category

6 Deficiencies and Improvements

As noted above, there are various issues with the analysis as completed so far. Below I discuss various issues and ways to address them that I believe will improve the analysis.

6.1 Increasing number of observations

The most important step is to increase the number of observations available, specifically the number of trials matched to ICD-10 codes with corresponding population estimates in the Global Burden of Disease Dataset. This will open up some important analysis options. Right now, the diseases are only being controlled for at very general levels. By increasing the number of observations it will allow for analysis across general categories, such as for different types of cancer. Matching this data is difficult to do by hand, but initial experiments suggest that Large Language Models may serve well as an automated process, making this data more accessible. Alternatively, this data may exists in a commercial data set.

6.2 Enrollment Modelling

One of the original goals of this project was to examine the impact that enrollment struggles have on the probability of trial termination. Unfortunately, this requires a model of clinical trial enrollment, and the data to estimate this data-set is woefully incomplete in my data set. If it were available, instead of the treatment being an instantaneous extension in the recruitment period, I could model the impact of slowing the enrollment rate.

There has been substantial work on forecasting multi-site enrollment rates and durations by [Toz+96; Car04; AF07; ZL10; ZL12a; ZL12b; HGY15; Jia+15; DZL17; LTH19; ZH22; USM22; BAM22; Ava+23] but choosing between the various single and multi-site models presented is difficult without a data set with which to validate the results and the data available at clinicaltrials.gov is incomplete. In most cases the trial sponsor reports the anticipated enrollment value while the trial is still recruiting and only updates the actual enrollment after the trial has ended. Some trials do publish an incremental record of their enrollment numbers, but this is not the norm. It may be possible to impute the enrollment process if a suitable model can be created.

6.3 Improving Measures of Market Conditions

In addition to the fact that many diseases may be treated by non-pharmaceutical means (e.g. diet, physical therapy, medical devices, etc), off-label prescription of pharmaceuticals is legal at the federal level [Com30]. These two facts both complicate measuring competing treatments, a key part of market conditions. One way to address non-pharmaceutical treatments is to concentrate on domains that are primarily treated by pharmaceuticals. Another way to address this would be to focus the analysis on just a few specific diseases, for which a history of treatment options can be compiled. Cancer treatments are a good example of both of these, because the treatments tend to be some combination of pharmacological products and radiation therapy or surgery. This would require identifying diseases that are prime candidates and then trials and treatments associated with those diseases.

Another area of focus could be in identifying diseases where the interaction between the number of drugs on the market and population are unlikely to be jointly determined. This might be possible if studying genetic diseases or endemic diseases.

7 Conclusion

The successful completion of Phase III clinical trials is crucial for bringing new treatments to market. This work establishes the first framework for separating the causal effects of operational versus strategic factors in clinical trial completion. Initial results indicate that enrollment period delays increase clinical trial termination probability by 2.5 percentage points. The approach developed here

can be extended with additional data on enrollment to provide more definitive guidance on the impact of enrollment delays. Further research in this direction could help reduce operational barriers to trial completion or estimating the impact policies may have through operational channels such as the effect they may have on enrollment. Identifying and removing impediments to completing clinical trials in otherwise capable pharmaceutical products has the potential to strengthen the drug development pipeline and improve the treatment options available to patients.

8 References

References

[22] *FDA Drug Approval Process*. Drugs.com. May 28, 2022. URL: https://www.drugs.com/fda-approval-process.html (visited on 04/12/2023).

[AAM04] Rosa M. Abrantes-Metz, Christopher Adams, and Albert D. Metz. "Pharmaceutical Development Phases: A Duration Analysis". In: *SSRN Electronic Journal* (2004). ISSN: 1556-5068. DOI: 10.2139/ssrn.607941. URL: http://www.ssrn.com/abstract=607941 (visited on 01/31/2023).

[AF07] Vladimir V. Anisimov and Valerii V. Fedorov. "Modelling, Prediction and Adaptive Adjustment of Recruitment in Multicentre Trials". In: *Statistics in Medicine* 26.27 (Nov. 30, 2007), pp. 4958–4975. ISSN: 02776715, 10970258. DOI: 10.1002/sim.2956. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.2956 (visited on 06/06/2023).

[Ava+23] Alejandra Avalos-Pacheco et al. "Validation of Predictive Analyses for Interim Decisions in Clinical Trials". In: *JCO precision oncology* 7 (Feb. 2023), e2200606. ISSN: 2473-4284. DOI: 10.1200/PO.22.00606. pmid: 36848613.

[BAM22] Cameron Bieganek, Constantin Aliferis, and Sisi Ma. "Prediction of Clinical Trial Enrollment Rates". In: *PLOS ONE* 17.2 (Feb. 24, 2022). Ed. by Sathishkumar V E, e0263193. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0263193. URL: https://dx.plos.org/10.1371/journal.pone.0263193 (visited on 05/02/2023).

[Car04] Rickey Edward Carter. "Application of Stochastic Processes to Participant Recruitment in Clinical Trials". In: *Controlled Clinical Trials* 25.5 (Oct. 2004), pp. 429–436. ISSN: 01972456. DOI: 10.1016/j.cct.2004.07.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0197245604000522 (visited on 04/27/2023).

[Cen22a] Centers For Medicare and Medicaid. *2022 ICD-10-PCS Official Guidelines for Coding and Reporting*. Apr. 2022.

[Cen22b] Centers For Medicare and Medicaid. *ICD-10-CM Official Guidelines for Coding and Reporting*. Apr. 2022.

[Cer07] Rodrigo A. Cerda. "Endogenous Innovations in the Pharmaceutical Industry". In: *Journal of Evolutionary Economics* 17.4 (June 27, 2007), pp. 473–515. ISSN: 0936-9937, 1432-1386. DOI: 10.1007/s00191-007-0059-3. URL: http://link.springer.com/10.1007/s00191-007-0059-3 (visited on 09/06/2024).

[Cli22] Clinical Trials Transformation Initiative (CTTI. *Aggregate Analysis of ClinicalTrials.Gov (AACT) Database*. https://aact.ctti-clinicaltrials.org, 2022.

[Com28] Office of the Commissioner. *The Drug Development Process*. FDA. Thu, 02/20/2020 - 17:28. URL: https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process (visited on 11/05/2024).

[Com30] Office of the Commissioner. *Understanding Unapproved Use of Approved Drugs "Off Label"*. FDA. Thu, 04/18/2019 - 00:30. URL: https://www.fda.gov/patients/learn-about-expanded-access-and-other-treatment-options/understanding-unapproved-use-approved-drugs-label (visited on 04/10/2023).

[DiM+10] J A DiMasi et al. "Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs". In: *Clinical Pharmacology & Therapeutics* 87.3 (Mar. 2010), pp. 272–277. ISSN: 0009-9236, 1532-6535. DOI: 10.1038/clpt.2009.295. URL: http://doi.wiley.com/10.1038/clpt.2009.295 (visited on 09/04/2024).

[DiM02] Joseph A. DiMasi. "The Value of Improving the Productivity of the Drug Development Process: Faster Times and Better Decisions". In: *PharmacoEconomics* 20 (Supplement 3 2002), pp. 1–10. ISSN: 1170-7690. DOI: 10.2165/00019053-200220003-00001. URL: http://link.springer.com/10.2165/00019053-200220003-00001 (visited on 10/11/2024).

[DZL17] Yi Deng, Xiaoxi Zhang, and Qi Long. "Bayesian Modeling and Prediction of Accrual in Multi-Regional Clinical Trials". In: *Statistical Methods in Medical Research* 26.2 (Apr. 2017), pp. 752–765. ISSN: 0962-2802, 1477-0334. DOI: 10.1177/0962280214557581. URL: http://journals.sagepub.com/doi/10.1177/0962280214557581 (visited on 04/27/2023).

[HGY15] Daniel F. Heitjan, Zhiyun Ge, and Gui-shuang Ying. "Real-Time Prediction of Clinical Trial Enrollment and Event Counts: A Review". In: *Contemporary Clinical Trials* 45 (Nov. 2015), pp. 26–33. ISSN: 15517144. DOI: 10.1016/j.cct.2015.07.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S1551714415300483 (visited on 05/02/2023).

[Hwa+16] Thomas J. Hwang et al. "Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results". In: *JAMA Internal Medicine* 176.12 (Dec. 1, 2016), p. 1826. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2016.6008. URL: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2016.6008 (visited on 01/31/2023).

[Jia+15] Yu Jiang et al. “Modeling and Validating Bayesian Accrual Models on Clinical Data and Simulations Using Adaptive Priors”. In: *Statistics in Medicine* 34.4 (Feb. 20, 2015), pp. 613–629. ISSN: 02776715. DOI: 10.1002/sim.6359. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.6359 (visited on 04/27/2023).

[Khm21] Ekaterina Khmelnitskaya. “Competition and Attrition in Drug Development”. University of Virginia, May 2021. 55 pp.

[LTH19] Yu Lan, Gong Tang, and Daniel F. Heitjan. “Statistical Modeling and Prediction of Clinical Trial Recruitment”. In: *Statistics in Medicine* 38.6 (Mar. 15, 2019), pp. 945–955. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8036. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.8036 (visited on 04/27/2023).

[Pea09] Judea Pearl. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge, U.K. ; New York: Cambridge University Press, 2009. 384 pp. ISBN: 978-0-521-89560-6 978-0-521-77362-1.

[Sta22] Stan Development Team. *Stan Modelling usersGuide and Reference Manual*. 2022. URL: https://mc-stan.org/.

[Sta23] Stan Development Team. *RStan: The R Interface to Stan*. 2023. URL: https://mc-stan.org/.

[Toz+96] A. E. Tozzi et al. “Predicting the Accrual Rate in a Vaccine Clinical Trial: An a Posteriori Evaluation of the Feasibility Study”. In: *Revue D'epidemiologie Et De Sante Publique* 44.5 (Oct. 1996), pp. 387–393. ISSN: 0398-7620. pmid: 8933665.

[US a] U.S. Food and Drug Administration. *Indexing Spl Fact Sheet*. U.S. Food and Drug Administration. URL: https://www.fda.gov/media/85645/download (visited on 04/08/2023).

[US b] U.S. National Library of Medicine. *FDAAA 801 and the Final Rule - ClinicalTrials.Gov*. ClinicalTrials.gov. URL: https://clinicaltrials.gov/ct2/manage-recs/fdaaa (visited on 04/08/2023).

[US c] U.S. National Library of Medicine. *Medical Subject Headings - Home Page*. URL: https://www.nlm.nih.gov/mesh/meshhome.html (visited on 04/09/2023).

[US d] U.S. National Library of Medicine. *RxNav-in-a-Box - RxNav Applications*. URL: https://lhncbc.nlm.nih.gov/RxNav/applications/RxNav-in-a-Box.html (visited on 04/10/2023).

[US e] U.S. National Library of Medicine. *RxNorm Overview*. URL: https://www.nlm.nih.gov/research/umls/rxnorm/overview.html (visited on 04/08/2023).

[US 24] U.S. National Library of Medicine. *Protocol Registration Data Element Definitions for Interventional and Observational Studies | ClinicalTrials.Gov*. ClinicalTrials.gov. June 17, 2024. URL: https://clinicaltrials.gov/policy/protocol-definitions (visited on 01/25/2025).

[USM22] Szymon Urbas, Chris Sherlock, and Paul Metcalfe. "Interim Recruitment Prediction for Multi-Center Clinical Trials". In: *Biostatistics* 23.2 (Apr. 13, 2022), pp. 485–506. ISSN: 1465-4644, 1468-4357. DOI: 10.1093/biostatistics/kxaa036. URL: https://academic.oup.com/biostatistics/article/23/2/485/5911853 (visited on 05/03/2023).

[Vos+20] Theo Vos et al. "Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019". In: *The Lancet* 396.10258 (Oct. 17, 2020), pp. 1204–1222. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)30925-9. pmid: 33069326. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30925-9/fulltext (visited on 03/14/2023).

[Wora] World Health Organization. *ICD-10 Version:2019*. ICD-10 Version:2019. URL: https://icd.who.int/browse10/2019/en (visited on 01/20/2025).

[Worb] World Health Organization. *International Classification of Diseases (ICD)*. URL: https://www.who.int/standards/classifications/classification-of-diseases (visited on 04/09/2023).

[ZH22] Xiaoxi Zhang and Bo Huang. "A Simple and Robust Model for Enrollment Projection in Clinical Trials". In: *Contemporary Clinical Trials* 123 (Dec. 2022), p. 106999. ISSN: 15517144. DOI: 10.1016/j.cct.2022.106999. URL: https://linkinghub.elsevier.com/retrieve/pii/S1551714422003251 (visited on 05/03/2023).

[ZL10] Xiaoxi Zhang and Qi Long. "Stochastic Modeling and Prediction for Accrual in Clinical Trials". In: *Statistics in Medicine* (2010), n/a–n/a. ISSN: 02776715, 10970258. DOI: 10.1002/sim.3847. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.3847 (visited on 04/27/2023).

[ZL12a] Xiaoxi Zhang and Qi Long. "Joint Monitoring and Prediction of Accrual and Event Times in Clinical Trials: Prediction of Accrual and Event Times in Clinical Trials". In: *Biometrical Journal* 54.6 (Nov. 2012), pp. 735–749. ISSN: 03233847. DOI: 10.1002/bimj.201100180. URL: https://onlinelibrary.wiley.com/doi/10.1002/bimj.201100180 (visited on 04/27/2023).

[ZL12b] Xiaoxi Zhang and Qi Long. "Modeling and Prediction of Subject Accrual and Event Times in Clinical Trials: A Systematic Review". In: *Clinical Trials* 9.6 (Dec. 2012), pp. 681–688. ISSN: 1740-7745, 1740-7753. DOI: 10.1177/1740774512447996. URL: http://journals.sagepub.com/doi/10.1177/1740774512447996 (visited on 04/27/2023).

A Other Statistical Results

Table 2: 5th Percentiles for overall distribution of differences

Percentile	Value
0%	-0.22566
5%	-0.01285
10%	-0.00521
15%	-0.00130
20%	-0.00005
25%	0.00003
30%	0.00048
35%	0.00215
40%	0.00480
45%	0.00772
50%	0.01079
55%	0.01411
60%	0.01777
65%	0.02189
70%	0.02659
75%	0.03207
80%	0.03857
85%	0.04655
90%	0.05706
95%	0.07317
100%	0.27211

Table 3: Table of summary statistics for Relative Fixed Effects | By ICD-10 Category

ICD-10 Category	sample size	mean	$P(\delta_p \geq 0)$	2.5%	5%	25%	50% median	75%	95%	97.5%
01 Infections & Parasites	20	0.135	0.772	-0.223	-0.163	0.013	0.134	0.257	0.438	0.498
02 Neoplasms	49	0.211	0.902	-0.107	-0.056	0.099	0.208	0.320	0.489	0.546
03 Blood & Immune system	1	0.061	0.624	-0.333	-0.267	-0.071	0.061	0.194	0.388	0.454
04 Endocrine, Nutritional, and Metabolic	10	0.177	0.820	-0.201	-0.137	0.045	0.173	0.304	0.505	0.573
05 Mental & Behavioral	8	0.142	0.767	-0.239	-0.175	0.010	0.140	0.270	0.469	0.537
06 Nervous System	8	0.009	0.524	-0.383	-0.315	-0.120	0.011	0.140	0.327	0.389
07 Eye and Adnexa	11	0.061	0.624	-0.333	-0.265	-0.069	0.062	0.192	0.386	0.452
08 Ear and Mastoid	0	0.063	0.625	-0.332	-0.263	-0.069	0.064	0.196	0.393	0.457
09 Circulatory	3	0.037	0.577	-0.362	-0.293	-0.094	0.038	0.170	0.363	0.426
10 Respiratory	7	0.061	0.622	-0.335	-0.269	-0.072	0.061	0.195	0.391	0.458
11 Digestive	12	0.060	0.621	-0.339	-0.270	-0.071	0.059	0.192	0.388	0.454
12 Skin & Subcutaneaous tissue	9	0.104	0.704	-0.283	-0.220	-0.027	0.103	0.234	0.429	0.494
13 Musculoskeletal	17	0.159	0.794	-0.218	-0.156	0.027	0.156	0.286	0.482	0.550
14 Genitourinary	2	0.063	0.627	-0.337	-0.270	-0.070	0.065	0.195	0.389	0.456
15 Pregancy, Childbirth, & Puerperium	0	0.061	0.622	-0.336	-0.270	-0.071	0.061	0.194	0.392	0.458
16 Perinatal Period	0	0.062	0.626	-0.338	-0.270	-0.068	0.062	0.194	0.390	0.457
17 Congential	3	0.063	0.628	-0.332	-0.264	-0.068	0.063	0.195	0.391	0.456
18 Symptoms, Signs etc.	1	0.062	0.625	-0.332	-0.265	-0.069	0.062	0.194	0.390	0.456
19 Injury etc.	0	0.063	0.625	-0.337	-0.268	-0.069	0.062	0.196	0.392	0.459
20 External Causes	0	0.063	0.625	-0.330	-0.266	-0.070	0.064	0.194	0.392	0.460
21 Contact with Healthcare	0	0.063	0.627	-0.335	-0.267	-0.070	0.064	0.194	0.390	0.454
22 Special Purposes	0	0.062	0.625	-0.335	-0.268	-0.070	0.062	0.194	0.391	0.456

Contents