Supplementary Content

## Supplementary Section 1. Example of raw data

We present an example of the raw Citeline data in Supplementary Supplementary Table 1. Each row consists of a unique drug and indication being test in a clinical trial. A trial can be repeated if the across the rows if it tests multiple drugs in its different arms.

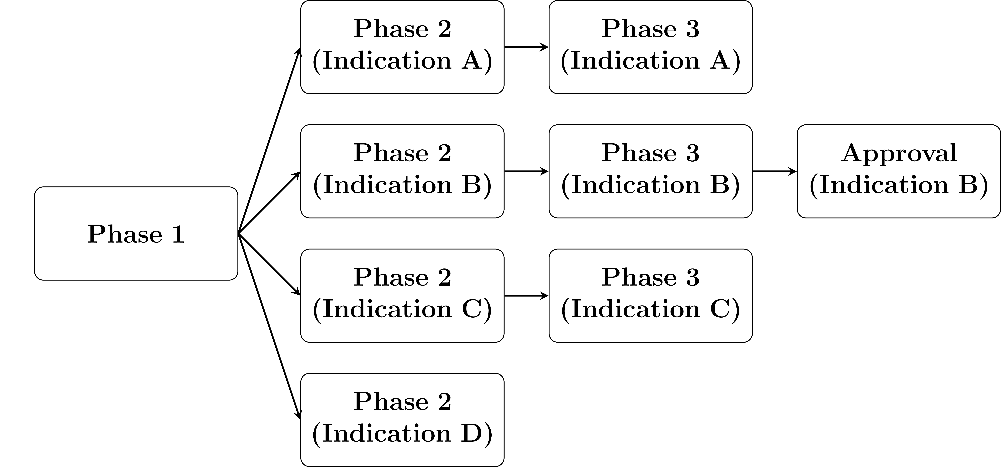
Supplementary Table 1. An example of the data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TrialID | Drug Name | Phase | Start Date | End Date | Disease Type | Therapeutic Area |
| 48391 | Loratadine | 1/2 | NULL | 7/6/2003 | Allergic Rhinitis | Autoimmune/ Inflammation |
| 70538 | Loratadine | 3 | NULL | 18/9/2007 | Allergic Rhinitis | Autoimmune/ Inflammation |
| 100378 | Loratadine | 3 | NULL | 29/10/2008 | Asthma | Autoimmune/ Inflammation |
| 122164 | Loratadine | 4 | 1/1/2010 | 1/3/2012 | Allergic Rhinitis | Autoimmune/ Inflammation |
| 151465 | Loratadine | 3 | 1/5/2011 | 14/5/2014 | Pain (nociceptive) | CNS |
| 153368 | Loratadine | 1 | NULL | 1/7/2006 | Asthma | Autoimmune/ Inflammation |

## Supplementary Section 2. Details of the model and computations

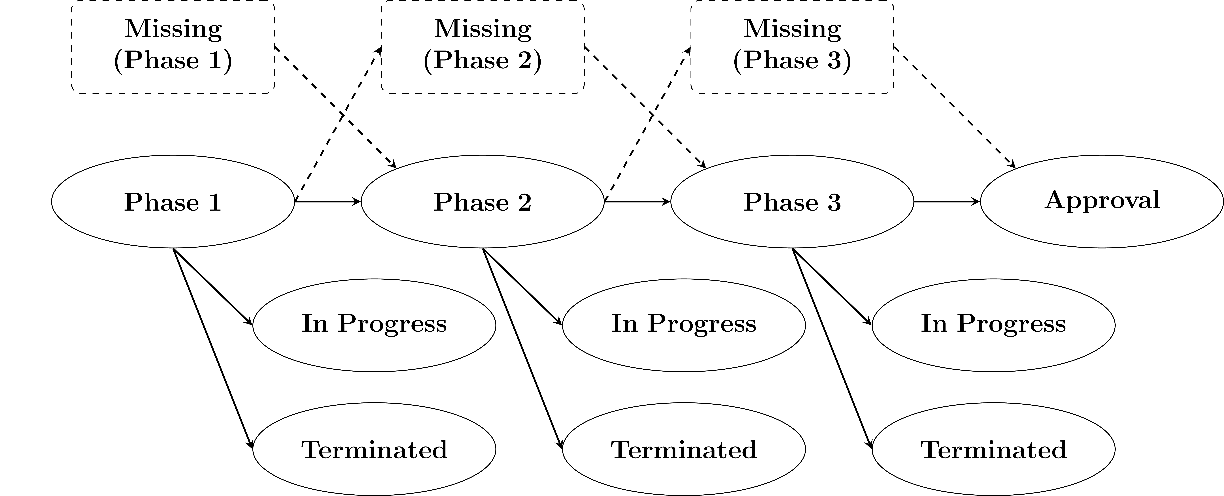
In this section, we elaborate on model of drug development and the detailed computations by replicating parts from Wong (2021).

We say that a drug development program has reached phase *i* if it is observed, or can be inferred, that there is at least one trial in phase *i*. It is possible that a clinical trial can be repeated in multiple development paths. In Supplementary Supplementary Figure 1, we show an example in which a single phase 1 trial for a drug is involved in four different development paths, each targeting a different disease. It is not uncommon that the result of the phase 1 trial is used as supporting evidence for the safe use of a drug, allowing that drug to be used for different indications without additional phase 1 testing. For example, hydroxychloroquine — already approved for the treatment of malaria — was tested for effectiveness against COVID-19 without another phase 1 clinical trial. There also exist clinical trials where different drug combinations are tested for the same indication in different arms.



Supplementary Figure 1

A *phase transition* between phase *i* and *j* is the change between the states of a drug development program. We make the assumption that each program must transition from phase 1 to phase 2 to phase 3 to approval in this order, and model the possible states in a drug development program as a Markov chain shown in Supplementary Supplementary Figure 2. Every drug development path in our study must start from phase 1 (or ‘missing’ phase 1) and end up in one of the nodes labeled as ‘in progress’, ‘terminated’ or ‘approval’.



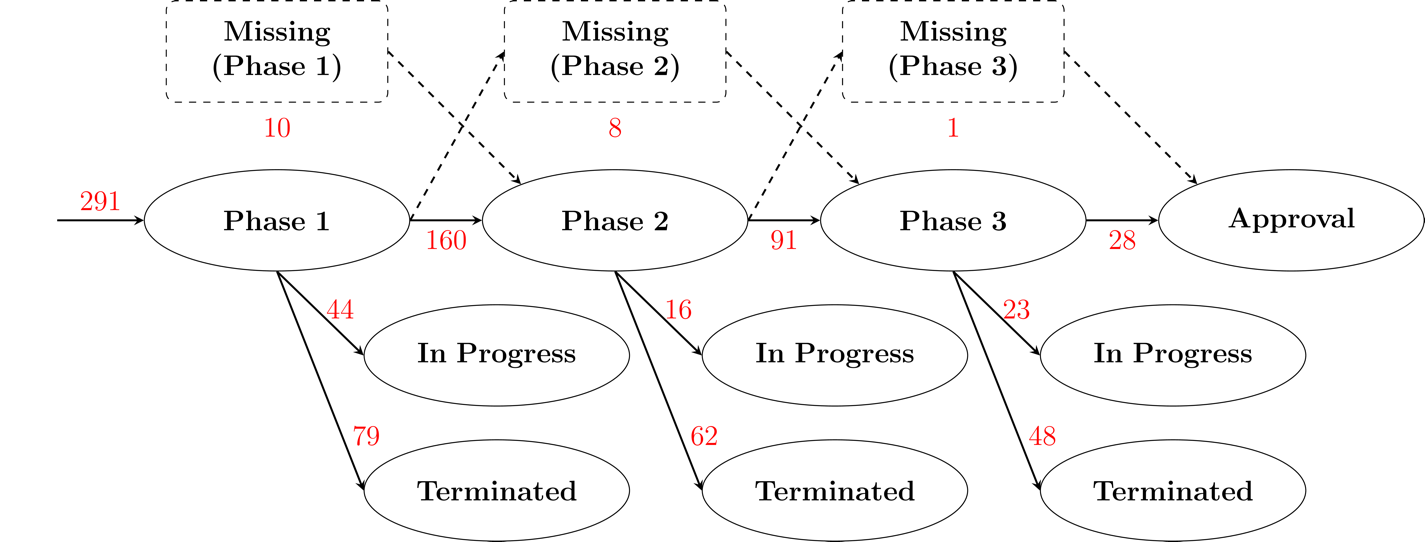
Supplementary Figure 2

We infer missing transitions in the development paths arising from incomplete records. This is plausible since each of these stages involves distinct predefined tests, all of which are required by regulators in any new drug application (NDA). If we observe data for phases 1 and 3 but not phase 2 trials for a given drug-indication pair, our idealized process implies that there was at least one phase 2 trial that occurred, but is missing from our dataset. Accordingly, we impute the successful completion of phase 2 in these cases. We make the standard assumption that phase 1/2 and phase 2/3 trials are to be considered as phase 2 and phase 3, respectively.

The probability of a drug development program transitioning from phase *i* to phase *j* (PoS*ij*) can be computed using the simple ratio Nj/Ni, where Ni is the number of drug development programs that have reached phase *i* (where *i* = 1, 2, 3) of the drug development process and are not in active development between phase *i* and phase *j* (where *j* = 2, 3, or “A” which denotes regulatory approval, *i* < *j*), and N*j* is the number of drug development programs among the former that made it to phase j. PoS1A is also known as the “overall PoS”.

We give some calculations using the example shown in Supplementary Figure 3

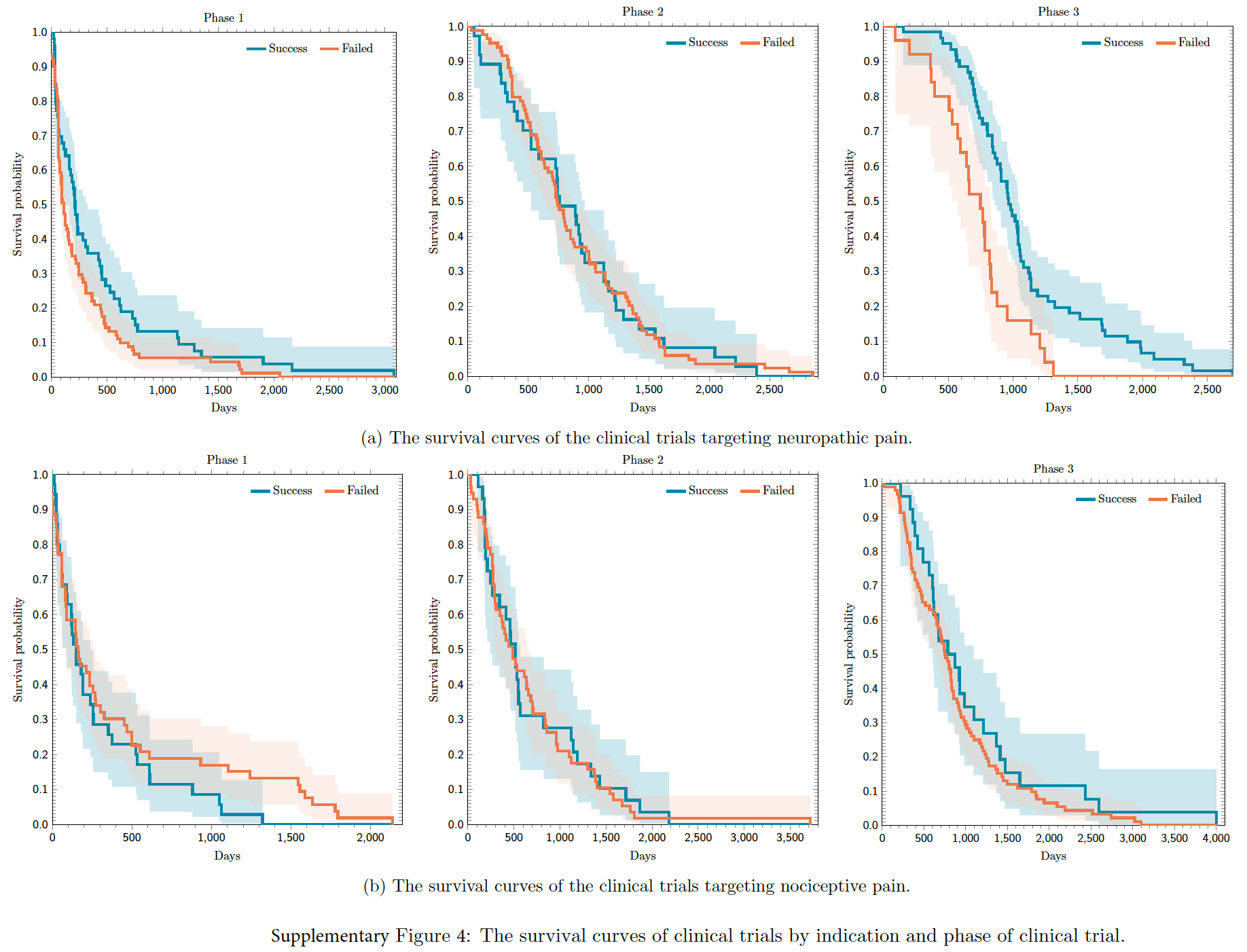
In that figure, we see that 291 development programs have conducted phase 1 testing whereas 10 programs have skipped phase 1 to go to phase 2 testing directly. This is not uncommon in many drug development programs, where drug candidates move directly to the higher phases based on safety studies conducted for other indications. Among these 301 drug development programs, we know that 44 have yet to conclude phase 1 testing while 257 have completed phase 1. Of these 257 programs, 178 have gone on to phase 2 while 79 have failed. In the notation introduced earlier, N1=257 and N2=178, yielding an estimate of 69.3% for PoS12. Repeating the logic for the transitions between phase 2 and phase 3, and between phase 3 and approval, gives 61.3% and 37.7% as estimates of PoS23 and PoS3A respectively.



Supplementary Figure 3

In order to compute the probability of a drug development program making it all the way from phase 1 to approval, we consider only the drug development programs that have definite outcomes. In other words, we do not consider development programs that are “in progress” in the denominator. In our example, the number of such programs is 10 + 291 - 44 - 16 - 23 = 218. Since 29 programs made it to approval, the estimated PoS1A is 29/218 = 13.3%.

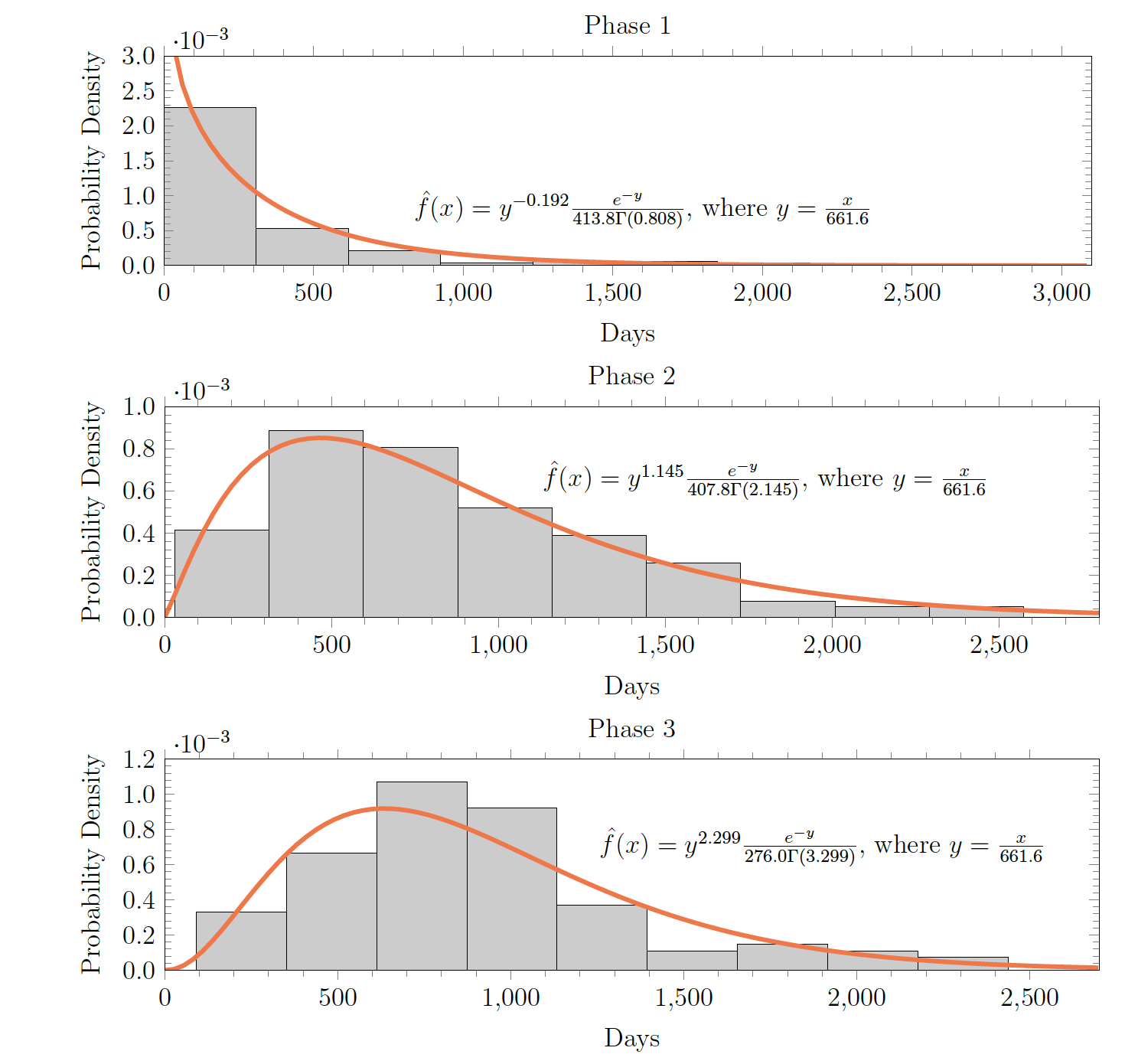
Supplementary Section 3: Survival curves of clinical trials



The survival curves presented in this supplementary figure are stratified by indication, and phase. The duration of a clinical trial is defined as the number of days from the initiation to the conclusion of subject enrollment or trial termination.

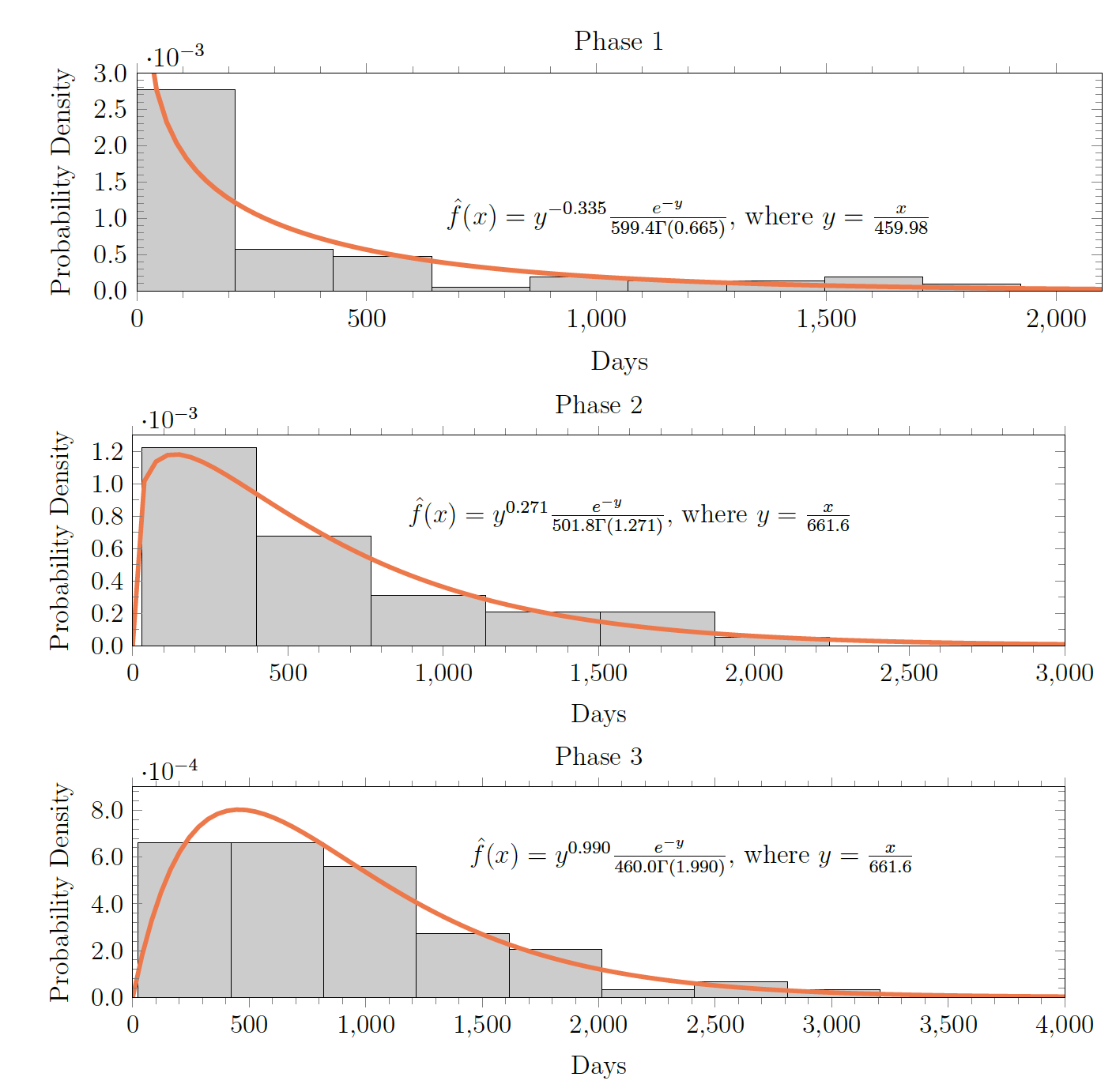
Supplementary Section 4: The distribution of the duration of clinical trials, together with the fitted gamma distributions.

Supplementary Figure 5.



Supplementary Figure 5: The distribution of the duration of clinical trials targeting neuropathic pain. We also model the density with a gamma distribution.

Supplementary Figure 6.



Supplementary Figure 6: The distribution of the duration of clinical trials targeting nociceptive pain. We also model the density with a gamma distribution.

# Supplementary Section 5: Logistic regression of drug attributes on outcome

We attempted to analyze the factors that contribute to the success of a clinical trial, defined as allowing the development program to progress to a higher phase, by regressing the features of the drug on the outcome with a logistic regression. Our feature set consists of the duration of the clinical trial, the pharmacology targets (e.g., opioid receptor or calcium channel, etc), therapy type (e.g., small molecule, macromolecule, or plant extract), whether the compounds involved have components with higher abuse potential, and the phase of the trial. Multiple compounds tested in a clinical trial have their individual properties encoded as a one-hot vector.

Our full feature set contains 51 dimensions, but the sparseness of some of the features prevented the estimation algorithm from converging. To resolve these issues, we removed features that did not meet a variance threshold of 0.09.

We estimated two models, one for clinical trials targeting neuropathic pain and the other for those targeting nociceptive pain. The results are shown in Supplementary Tables 2 and 3.

As can be seen, the properties of the drugs involved in the clinical trials are poor predictors of whether the clinical trial will move the development program forward.

Supplementary Table 2: Logit regression results for clinical trials targeting neuropathic pain

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. Observations |  | 200 |  | | |
| Degree of freedom |  | 9 |
| Pseudo R-squared |  | 0.09606 |
| Log-Likelihood |  | -106.33 |
| LL-Null |  | -117.63 |
| LLR p-value |  | 0.007158 |
| Variable  Duration |  | Coefficient  0.00 | Std. Err  0.00 | z  1.42 | *P > |z|*  0.16 |
| Pharmacology Grp: | opioid receptor | -0.19 | 0.64 | -0.30 | 0.76 |
| Pharmacology Grp: | other | -0.33 | 0.46 | -0.72 | 0.47 |
| Pharmacology Grp: | sodium channel | 0.34 | 0.50 | 0.70 | 0.49 |
| Pharmacology Grp: | calcium channel | 0.67 | 0.63 | 1.05 | 0.29 |
| Type: small molecule | | -0.50 | 0.57 | -0.88 | 0.38 |
| High abuse potential: Yes | | 0.81 | 0.51 | 1.58 | 0.11 |
| Phase: 1 | | -0.76 | 0.60 | -1.28 | 0.20 |
| Phase: 2 | | -1.71 | 0.63 | -2.72 | 0.01 |
| Phase: 3 | | -1.07 | 0.76 | -1.41 | 0.16 |

Supplementary Table 3: Logit regression results for clinical trials targeting nociceptive pain.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. Observations |  | 227 |  |  |  |
| Degree of freedom |  | 10 |  |  |  |
| Pseudo R-squared |  | 0.03146 |  |  |  |
| Log-Likelihood |  | -137.35 |  |  |  |
| LL-Null |  | -141.81 |  |  |  |
| LLR p-value |  | 0.5395 |  |  |  |
| Variable  Duration |  | Coefficient  0.00 | Std. Err  0.00 | z  1.16 | *P > |z|*  0.25 |
| Pharmacology Grp: | opioid receptor | 0.91 | 0.61 | 1.48 | 0.14 |
| Pharmacology Grp: | other | -0.43 | 0.38 | -1.14 | 0.25 |
| Pharmacology Grp: | cytokine inhibitor | 0.10 | 0.58 | 0.18 | 0.86 |
| Pharmacology Grp: | sodium channel | 0.23 | 0.42 | 0.54 | 0.59 |
| Type: small molecule | | 0.59 | 0.84 | 0.71 | 0.48 |
| Type: biologic | | 0.43 | 0.93 | 0.46 | 0.65 |
| High abuse potential: Yes | | -0.35 | 0.55 | -0.63 | 0.53 |
| Phase: 1 | | -1.51 | 0.88 | -1.71 | 0.09 |
| Phase: 2 | | -1.52 | 0.88 | -1.73 | 0.08 |
| Phase: 3 | | -1.63 | 0.85 | -1.91 | 0.06 |