Biostatistics (2018) **00**, 00, *pp.* 1–14
doi:10.1093/biostatistics/kxx069

Estimation of clinical trial success rates and related parameters

CHI HEEM WONG, KIEN WEI SIAH

MIT Computer Science and Artificial Intelligence Laboratory & Department of Electrical Engineering and Computer Science, Cambridge, MA 02139, USA and MIT Sloan School of Management and Laboratory for Financial Engineering, Cambridge, MA 02142, USA

ANDREW W. LO*

MIT Computer Science and Artificial Intelligence Laboratory & Department of Electrical Engineering and Computer Science, Cambridge, MA 02139, USA, MIT Sloan School of Management and Laboratory for Financial Engineering, Cambridge, MA 02142, USA, and AlphaSimplex Group, LLC, Cambridge, MA 02142, USA

alo-admin@mit.edu

Summary

Previous estimates of drug development success rates rely on relatively small samples from databases curated by the pharmaceutical industry and are subject to potential selection biases. Using a sample of 406 038 entries of clinical trial data for over 21 143 compounds from January 1, 2000 to October 31, 2015, we estimate aggregate clinical trial success rates and durations. We also compute disaggregated estimates across several trial features including disease type, clinical phase, industry or academic sponsor, biomarker presence, lead indication status, and time. In several cases, our results differ significantly in detail from widely cited statistics. For example, oncology has a 3.4% success rate in our sample vs. 5.1% in prior studies. However, after declining to 1.7% in 2012, this rate has improved to 2.5% and 8.3% in 2014 and 2015, respectively. In addition, trials that use biomarkers in patient-selection have higher overall success probabilities than trials without biomarkers.

Keywords: Clinical phase transition probabilities; Clinical trial statistics; Probabilities of success.

1. Introduction

The probability of success (POS) of a clinical trial is critical for clinical researchers and biopharma investors to evaluate when making scientific and economic decisions. Prudent resource allocation relies on the accurate and timely assessment of risk. Without up-to-date estimates of the POS, however, investors may misjudge the risk and value of drug development, leading to lost opportunities for both investors and patients.

One of the biggest challenges in estimating the success rate of clinical trials is access to accurate information on trial characteristics and outcomes. Gathering such data is expensive, time-consuming, and

*To whom correspondence should be addressed.

susceptible to error. Previous studies of success rates have been constrained by the data in several respects. Abrantes-Metz *and others* (2005) surveyed 2328 drugs using 3136 phase transitions (e.g., from Phase 1 to Phase 2 in the approval process), while DiMasi *and others* (2010) studied 1316 drugs from just 50 companies. In the landmark study of this area, Hay *and others* (2014) analyzed 7372 development paths of 4451 drugs using 5820 phase transitions. In two recent papers, Smietana *and others* (2016) computed statistics using 17 358 phase transitions for 9200 compounds, while Thomas *and others* (2016) used 9985 phase transitions for 7455 clinical drug development programs. In contrast, ClinicalTrials.gov, the clinical trial repository maintained by the National Institutes of Health (NIH), contains over 217 000 clinical trial entries submitted by various organizations as of July 1, 2016 (see www.clinicaltrials.gov). It is estimated that trained analysts would require tens of thousands of hours of labor to incorporate its full information manually to produce POS estimates.

In this article, we construct estimates of the POS and other related risk characteristics of clinical trials using 406 038 entries of industry- and non-industry-sponsored trials, corresponding to 185 994 unique trials over 21 143 compounds from Informa Pharma Intelligence's Trialtrove and Pharmaprojects databases from January 1, 2000 to October 31, 2015. This is the largest investigation thus far into clinical trial success rates and related parameters. To process this large amount of data, we develop an automated algorithm that traces the path of drug development, infers the phase transitions, and computes the POS statistics in hours. In this article, we introduce the "path-by-path" approach that traces the proportion of development paths that make it from one phase to the next. In contrast, extant literature uses what we call the "phase-by-phase" approach, which estimates the POS from a random sample of observed phase transitions. Apart from the gains in efficiency, our algorithmic approach allows us to perform previously infeasible computations, such as generating time-series estimates of POS and related parameters.

We estimate aggregate success rates, completion rates (CRs), phase-transition probabilities, and trial durations, as well as more disaggregated measures across various dimensions such as clinical phase, disease, type of organization, and whether biomarkers are used. Before presenting these and other results, we begin by discussing our methodology and describing some features of our data set.

2. DATA

We use Citeline data provided by Informa Pharma Intelligence, a superset of the most commonly used data sources that combines individual clinical trial information from Trialtrove and drug approval data from Pharmaprojects. In addition to incorporating multiple data streams, including nightly feeds from official sources such as ClinicalTrials.gov, Citeline contains data from primary sources such as institutional press releases, financial reports, study reports, and drug marketing label applications, and secondary sources such as analyst reports by consulting companies. Secondary sources are particularly important for reducing potential biases that may arise from the tendency of organizations to report only successful trials, especially those prior to the FDA Amendments Act of 2007, which requires all clinical trials to be registered and tracked via ClinicalTrials.gov. Our database contains information from both US and non-US sources.

The database encodes each unique quartet of trial identification number, drug, indication, and sponsor as a data point. As such, a single trial can be repeated as multiple data points. The trials range from January 1, 2000, to October 31, 2015, the latter being the date that we received the data set. After deleting 46 524 entries with missing dates and unidentified sponsors, and 1818 entries that ended before January 1, 2000, 406 038 data points remain. Of these, 34.7% (141 086) are industry sponsored and 65.3% (264 952) are non-industry sponsored. In our industry-sponsored analysis, we counted 41 040 development paths or 67 752 phase transitions after the imputation process. Figure S1 in Section A1 of the supplementary material available at *Biostatistics* online contains an illustrative sample of the data set and some basic summary information.

Some trials are missing end-dates due to the failure of their sponsors to report this information. Since these dates are required by our algorithm, we estimate them by assuming that trials lasted the median duration of all other trials with similar features. Only 14.6% (59 208) of the data points required the estimation of end-dates.

3. Modeling the drug development process

To avoid confusion and facilitate the comparison of our results with those in the extant literature, we begin by defining several key terms. A drug development program is the investigation of a particular drug for a single indication (see top diagram of Figure S2 of the supplementary material available at *Biostatistics* online). A drug development program is said to be in Phase i if it has at least one Phase i clinical trial. If a Phase i clinical trial concludes and its objectives are met, this trial is said to be completed. If it is terminated prematurely for any reason, except in the case that it has positive results, the trial is categorized as failed. Conditioned on one or more trial(s) being completed, the sponsor can choose to either pursue Phase $i+1$ trials, or simply terminate development. If the company chooses the former option, the drug development program is categorized as a success in Phase i, otherwise, it will be categorized as terminated in Phase i. See Figure S2 (bottom) of the supplementary material available at *Biostatistics* online for an illustration. The POS for a given Phase i, denoted by $\text{POS}_{i,i+1}$, is defined as the probability that the drug development program advances to the next phase. The probability of getting a drug development program in Phase i through to approval is denoted by $\text{POS}_{i,\text{APP}}$. Hence the overall probability of success—moving a drug from Phase 1 to approval, which Hay *and others* (2014) calls the likelihood of approval (LOA)—is $\text{POS}_{1,\text{APP}}$.

The proper interpretation of drug development programs from clinical trial data requires some understanding of the drug development process, especially in cases of missing data. This is particularly important for estimating a drug candidate's $\text{POS}_{1,\text{APP}}$, which is typically estimated by multiplying the empirical POS of Phase 1 (safety), 2 (efficacy for a given indication), and 3 (efficacy for larger populations and against alternatives) trials. If, for example, Phase 2 data are missing for certain approved drugs, the estimated $\text{POS}_{1,\text{APP}}$ would be biased downward. Here, we take a different approach to estimating POSs.

Consider an idealized process in which every drug development program passes through Phase 1, 2, and 3 trials, in this order. 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 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 data set. Accordingly, we impute the successful completion of Phase 2 in these cases. There exist some cases where Phase 2 trials are skipped, as with the recent example of Aducanumab (BIIB037), Biogen's Alzheimer's candidate, as reported by Root (2014). Since skipping Phase 2 trials is motivated by compelling Phase 1 data, imputing the successful completion of Phase 2 trials in these cases to trace drug development paths may not be a bad approximation. In addition, 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.

These assumptions allow us to more accurately reconstruct 'drug development paths' for individual drug-indication pairs, which in turn yield more accurate POS estimates. Let n^j be the number of drug development paths with observed Phase j trials, and n^j_s be the number of drug development paths where we observe phase transitions of state s of Phase j (defined below).

$$s = \begin{cases} ip, & \text{if all the trials are in progress} \\ t, & \text{if the program failed to proceed to phase } i+1 \text{ (i.e., terminated)} \\ m, & \text{if the phase transition can be inferred to be missing} \end{cases}$$

Equation 3.1 is the conservation law for drug development paths in Phase $j+1$.

$$n^{j+1} = n^j + n_m^j - n_{ip}^j - n_t^j \qquad \forall j = 1, 2, 3 \tag{3.1}$$

The POS from any one state to the next, $\text{POS}_{j,j+1}$, is thus the ratio of the number of drug development projects in Phase $j+1$, both observed and non-observed, to the number of drug development projects in Phase j, both observed and non-observed:

$$\text{POS}_{j,j+1}(\text{Path-by-Path}) = \frac{n^{j+1}}{n^j + n_m^j - n_{ip}^j} \tag{3.2}$$

Given our model, we can now compute $\text{POS}_{1,\text{APP}}$ by finding the proportion of development paths that made it from Phase 1 to Approval:

$$\text{POS}_{1,\text{APP}}(\text{Path-by-Path}) = \frac{n^{\text{Approval}}}{n^1 + n_m^1 - n_{ip}^1 - n_{ip}^2 - n_{ip}^3} \tag{3.3}$$

We term this the 'path-by-path' approach. In contrast, extant papers define the phase transition probability as the ratio of observed phase transitions to the number of observed drug development programs in Phase i and multiply the individual phase probabilities to estimate the overall POS. We term this the 'phase-by-phase' approach, which we shall differentiate from the path-by-path computation by a superscript p as follows:

$$\text{POS}_{j,j+1}^p = \frac{n^{j+1} - n_m^j}{n^j - n_{ip}^j} \tag{3.4}$$

$$\text{POS}_{1,\text{APP}}^p = \prod_{j \in \{1,2,3\}} \text{POS}_{j,j+1}^p \tag{3.5}$$

Implicit in the path-by-path computation method is the assumption that we have relatively complete information about the trials involved in drug development programs. This is true of our data set, as we are analyzing relatively recent years where trial pre-registration is a prerequisite for publication in major medical journals and use of the studies as supporting evidence for drug applications.

However, this assumption breaks down when we look at short windows of duration, for example, in a rolling window analysis to estimate the change in the POS over time. In such cases, we default back to the 'phase-by-phase' estimation to get an insight into the trend. This is done by considering only those drug development programs with phases that ended between t_1 and t_2 in the computation of the POS.

$$\text{POS}_{j,j+1}^p(t_1, t_2) = \frac{n^{j+1}(t_1, t_2) - n_m^j(t_1, t_2)}{n^j(t_1, t_2) - n_{ip}^j(t_1, t_2)} \tag{3.6}$$

$$\text{POS}_{1,\text{APP}}^p(t_1, t_2) = \prod_{j \in \{1,2,3\}} \text{POS}_{j,j+1}^p(t_1, t_2) \tag{3.7}$$

We further note that if no phase transitions are missing, the path-by-path and phase-by-phase methods should produce the same results, but the former will be more representative of actual approval rates if

phase transitions are missing. We elaborate on this in Section A2 of the supplementary material available at *Biostatistics* online.

Given our development-path framework, we can compute the POS using an algorithm that recursively considers all possible drug-indication pairs and determines the maximum observed phase. Reaching Phase i would imply that all lower phases were completed. To determine if a drug development program has been terminated in the last observed phase or is still ongoing, we use a simple heuristic: if the time elapsed between the end date of the most recent Phase i and the end of our sample exceeds a certain threshold t_i, we conclude that the trial has terminated. Based on practical considerations, we set t_i to be 360, 540, and 900 days for Phases 1, 2, and 3, respectively. For example, we assume that it takes approximately 6 months to prepare documents for an NDA filing after a Phase 3 trial has been completed. Since the FDA has a 6-month period to decide if it wishes to follow-up on a filing, and an additional 18 months to deliver a verdict, this places the overall time between Phase 3 and Approval to about 30 months, hence we set $t_3 = 900$ days. A pseudo-code for the algorithm is given in Figure S5 in Section A3 of the supplementary material available at *Biostatistics* online.

In summary, our algorithm allows us to impute missing trial data, and by counting the number of phase transitions, we can estimate the phase and overall POS.

4. Results

4.1. *POS for all drugs and indications*

Table 1 contains our estimates of the aggregate POS for each clinical phase across all indications. Corresponding estimates from the prior literature are also included for comparison. We find that 13.8% of all drug development programs eventually lead to approval, which is higher than the 10.4% reported by Hay *and others* (2014) and the 9.6% reported by Thomas *and others* (2016). The overall POS presented in this study, Hay *and others* (2014), and Thomas *and others* (2016) are much higher than the 1% to 3% that is colloquially seen as it is conditioned on the drug development program entering Phase 1. Our phase-specific POS estimates are higher in all phases. The largest increase is seen in $\text{POS}_{2,3}$, where we obtained a value of 58.3% compared to 32.4% in Hay *and others* (2014) and 30.7% in Thomas *and others* (2016). These differences may be due to our method of imputing missing clinical trials.

Table 2 contains phase and overall POS estimates by therapeutic group. The overall POS ($\text{POS}_{1,\text{APP}}$) ranges from a minimum of 3.4% for oncology to a maximum of 33.4% for vaccines (infectious disease). The overall POS for oncology drug development programs is about two-thirds the previously reported estimates of 5.1% in Thomas *and others* (2016) and 6.7% in Hay *and others* (2014).

A significantly different pattern emerges when we consider the phase POS for lead indications. The overall POS ($\text{POS}_{1,\text{APP}}$) increases when considering only lead indications, which is in line with the findings by Hay *and others* (2014). However, while we find an increase in the POS for Phase 1 ($\text{POS}_{1,2}$) and Phase 3 ($\text{POS}_{3,\text{APP}}$), we find a decrease in the POS for Phase 2 ($\text{POS}_{2,3}$) when looking only at lead indications. The POS for lead indications may be lower than the POS for all indications if a company initiates clinical trials for many indications, and most of them move on to the next phase. Conversely, the POS for lead indications may be higher if many of the initiated clinical trials for the same drug fail. The practice of initiating clinical trials for multiple indications using the same drug is prevalent in the industry, as documented in Table S2 in Section A5 of the supplementary material available at *Biostatistics* online. The relative performance of the various therapeutic groups remains the same when considering only lead indications, with oncology remaining the lowest performing group at 11.4% for $\text{POS}_{1,\text{APP}}$. Finally, the overall POS for individual therapeutic groups when considering only lead indications shows mixed directions in comparison to the respective overall POS specific to the indication.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.