Financing Innovation with Correlated Projects: Evidence from Drug Development*

Henry Conter† **Mihai Ion**‡

Abstract

Using a comprehensive, project-level database on drug development programs, we examine the link between external financing markets and drug approval outcomes. We document a significant increase in the percentage of projects explicitly citing lack of funding as the reason for termination and show that these terminations correlate strongly with periods of rising equity financing costs. Furthermore, we show evidence of growing clustering in drug development programs with respect to the biological targets they pursue, leading to more correlated approval outcomes. We examine whether this growing correlation has hampered the effectiveness of securitization techniques to mitigate the increasing underinvestment risk we document in this study. We show that a failure to diversify portfolios of drug development projects across biological targets significantly lowers their performance.

*This paper has benefited from comments and discussions with seminar participants at the University of Arizona, Arizona State University and University of Oklahoma.

†F. Hoffmann-La Roche AG, author email: henry.conter@roche.com

‡University of Arizona, author email: mihaiion@arizona.edu

1 Introduction

The process of developing and bringing a new drug to market is complex, costly, and fraught with uncertainty. Over the past several decades, pharmaceutical innovation has seen remarkable scientific advances, generating substantial social and economic welfare gains (Murphy and Topel (2006)). However, these advances are accompanied by ever-changing regulatory requirements, financial constraints, and shifting patterns of sponsor involvement. This paper examines the evolving landscape of drug development, the critical financial constraints faced by sponsors, and how changes in drug development concentration may impact both investment decisions and societal outcomes.

Drug development can be characterized as a multi-stage process involving substantial regulatory oversight, beginning with preclinical research and continuing through clinical trials, approval, and post-marketing surveillance. The development pathway requires sponsors – who range from very large pharmaceutical firms to small biotechnology startups – to invest considerable resources over an extended period, often without any guarantee of success. Each phase of clinical trials is designed to rigorously assess safety and efficacy before advancing to subsequent phases. This process is not only scientifically intensive but also a significant financial burden for many sponsors, particularly given the long gestation periods, high costs, and low probabilities of approval that characterize much of the industry (Lo and Thakor (2022)).

Our study draws on a rich dataset compiled from two proprietary databases provided by Citeline: Trialtrove, which tracks global clinical trials, and Pharmaprojects, which provides detailed information on drug development pipelines. Trialtrove provides over 170 pieces of information for each clinical trial, including data on trial design, sponsors, outcomes, and patient demographics. Pharmaprojects, meanwhile, offers comprehensive coverage of drugs from preclinical research through to post-marketing phases, with data on over 100 drug-level characteristics, such as molecular structure, mechanism of action, biological targets, and commercial status. The combined dataset includes information on approximately 400,000 clinical trials across 200 countries and over 100,000 unique drugs, with coverage spanning the past three decades. These databases provide a unique opportunity to examine the dynamics of drug development at the individual-project level. This allows us to investigate key characteristics such as scientific risk, phase transition probabilities, trial durations, and sponsor types, that play a critical role in shaping the financial risks and

rewards associated with drug development.

Despite the continuous innovations in medical research and an increasing understanding of disease mechanisms, the approval of a new drug remains an infrequent event. In our sample, the overall probability of success (PoS) from the preclinical stage to approval is only 8.3%, and, on average, only 20.8% of drugs that enter phase 1 are ultimately approved. These estimates are significantly lower for drugs targeting complex diseases such as cancer, where we observe a PoS from preclinical to approval of only 3.7% and a PoS from phase 1 to approval of only 8.4%. This low probability of success is one of the key challenges for the pharmaceutical industry and raises questions about the sustainability of current models of funding and innovation. Another distinctive feature of drug development is its prolonged timeline. Our data shows that, on average, it takes 6.05 years for a drug to progress from phase 1 to approval, with substantial variation across therapeutic areas (e.g. oncology drugs have an average development time of 10.12 years). These long development timelines delay the realization of revenues, increase the overall cost of drug development, and amplify the risk that sponsors will need to secure additional financing.

The financial risks associated with drug development are not uniformly distributed across the industry. Large pharmaceutical companies with established revenue streams from approved drugs are more capable of bearing the costs of development and the inherent risks of clinical failure. However, our data shows that the clinical trial process is dominated by small sponsors: the "Top 20 Pharma" companies (by size) sponsor only 20% of the clinical trials in our sample, while 56% of trials have academic sponsors, and 23% of trials are sponsored by smaller pharmaceutical companies. Given their limited sources of internally generated funds, these smaller sponsors are heavily reliant on external financing.[1] This creates a vulnerability in the drug development ecosystem, where projects of high societal value may be abandoned for reasons unrelated to their clinical potential, but rather due to financial constraints.[2] This underinvestment risk is further exacerbated by the technical complexity of drug development which creates strong information asymmetries between

[1]Recent evidence by Thakor et al. (2017) and Aghamolla and Thakor (2022) shows that the cost of capital for smaller biopharma firms, measured by their stock market betas, is significantly higher than for large, established pharmaceutical companies, reflecting their heightened exposure to financing risk.

[2]See Hall and Lerner (2010) and Kerr and Nanda (2015) for a more detailed discussion of this funding gap for R&D investment relative to the social optimum. As examples of this funding gap in drug development, Kyle (2018) shows evidence that countries are underinvesting in drugs with high therapeutic value (as measured by the French Haute Authorité de Santé) and Krieger, Li, and Papanikolaou (2022) show that firms increase their investment in novel drugs in response to a shock that reduced financial constraints (Medicare Part D).

suppliers of capital and drug developers.[3] Moreover, the low probabilities of approval make agency problems harder to detect, as it is difficult to ascertain whether project failures are due to moral hazard or the inherently low likelihood of success (see Thakor and Lo (2017) and Jorring et al. (2021)). These financing frictions can distort investment away from the first-best choice, causing firms to pass up on investments that may have not only significant societal value but also a positive net present value.

In the first part of our paper, we present evidence that this risk of underinvestment in drug development has increased significantly over the past two decades. First, we find that large, "Top 20 Pharma" companies are increasingly reducing their involvement in the clinical trial process, particularly in the early phases of development. The percentage of phase 1 trials sponsored by these large firms has declined from 28% in 2003 to just 11% in 2023. In contrast, smaller pharmaceutical companies and academic institutions, which are generally more financially constrained, are assuming a larger role in clinical research. Second, we show that clinical trial durations have increased substantially over time. Drug-development projects which concluded in 2023 took, on average, 40% longer to complete than projects which concluded in 2001. Third, we document a significant increase in the frequency of trials investigating drugs with unproven targets, which makes them more speculative and therefore riskier in terms of both clinical and commercial outcomes. For example, we find that, for clinical trials starting in 2023, 65% of drugs investigated had biological targets which had no prior approval. For trials starting in 1993, this number was only 32%.

The larger preponderance of small sponsors, significantly longer lags between cash outflows and inflows, and higher prevalence of unproven drug targets, all point to an increased exposure of drug development projects to fluctuations in the supply of external financing. Consistent with these trends we document a growing incidence of trial terminations due to lack of funding – from to 2.2% of all terminations in 2000 to 7.8% in 2023.[4] Furthermore, a regression analysis shows that, even after we control for this trend, funding-related terminations are significantly more likely during periods of low equity market sentiment, low IPO volume, or high equity market volatility.[5]

[3]This is the classic argument from Myers and Majluf (1984) who argue that asymmetric information leads to higher costs of external financing due to adverse selection.

[4]For these tests, we use data compiled by Citeline analysts on the reason why clinical trials were terminated.

[5]We measure equity market sentiment using the Baker and Wurgler (2006) index and equity market volatility using the VIX index from the CBOE. We focus on equity financing because biopharma firms make very little use of debt financing (e.g. Giambona et al. (2021)). The heavy reliance on equity financing is common for R&D intensive firms (Brown et al. (2009)) and is generally understood to be a consequence of their limited tangible assets to offer as

In light of these challenges, in the second part of our paper we investigate a securitization approach to financing drug development projects which may offer a solution to the increasing underinvestment risks outlined above. Securitization involves pooling a portfolio of drug development projects into a single investment vehicle, which can then issue different tranches of securities to investors with varying risk preferences. The underlying hypothesis is that, by aggregating multiple drug development projects, the idiosyncratic risks associated with individual projects—such as the failure of a specific clinical trial—can be diversified away, resulting in a more stable and attractive investment proposition. This concept, first introduced by Fernandez et al. (2012), highlights the potential for financial engineering to provide a more sustainable model for funding pharmaceutical innovation.

However, a critical consideration in the application of securitization to drug development is the degree of correlation between the individual projects within a portfolio. Correlations can arise when multiple drugs target the same biological pathway or mechanism of action, thereby increasing the likelihood that failures are not independent events. For example, if the underlying target is invalid or if the safety profile of the target is problematic, all drugs targeting that same pathway are likely to fail. Our analysis reveals a significant degree of concentration of drug development efforts around a limited number of biological targets. In our sample, we have 47,546 unique drugs with target data, and only 3,464 unique targets, which amounts to an average of 13.7 drugs per target. Moreover, we show that this concentration has increased substantially over time. The number of drugs investigated has significantly outpaced the number of targets (though both have increased steadily over time). Furthermore, the Gini index of inequality in target usage has increased from 0.42 in 1998 to 0.6 in 2021. Finally, we document a strong positive relation between a drug's approval probability and the average approval rate of drugs with the same target. This provides evidence of a strong link between target concentration and cross-drug correlations in approval outcomes.

Our investigation of the dynamics of drug-target concentrations reveals two opposing forces that may influence the extent to which portfolios of drug-development programs can achieve better risk-return tradeoffs. On one hand, the increased concentration of drugs and trials around similar targets

collateral (e.g. Rampini and Viswanathan (2013)), and high information asymmetry between borrowers and lenders (e.g. Besanko and Thakor (1987)). We use the IPO volume as a measure of the availability of equity financing, as it is a common way for biotechnology firms to raise capital (Aghamolla and Thakor (2022)).

means that, in recent periods, randomly selecting drugs for a portfolio is more likely to include drugs sharing the same target. Based on our results, this should lead to more correlated outcomes among the drugs in the portfolio. On the other hand, the larger number of available targets in recent periods presents an opportunity to enhance portfolio diversification by carefully selecting drugs with unrelated targets. We explore how these competing forces impact drug-portfolio performance through a series of simulations designed to assess how portfolio risks and average returns change with portfolio size. Specifically, we construct portfolios ranging from 1 to 100 drugs by randomly drawing drug-development projects from our Pharmaprojects dataset and recording their *realized* approval outcomes. Portfolio returns are measured as the number of drugs in the portfolio that achieved regulatory approval divided by the total clinical trial costs incurred.[6] This measure aims to capture societal value without having to assign relative weights to different medical conditions.

Our findings reveal significant benefits of diversification in drug development. Portfolios consisting of 20 drugs demonstrated a substantial decrease in risk compared to single-drug portfolios. Specifically, the standard deviation of returns decreased by 70%, and the likelihood of obtaining no approvals dropped from 79% to less than 1%. Additionally, the average returns nearly doubled when increasing the portfolio size from 1 to 20 drugs (from 0.57 approvals per billion spent, to 1.05). We find that expanding the portfolio beyond 20 drugs yielded minimal additional benefits in average returns, although it continued to reduce portfolio risk significantly.

To examine how increased correlations in drug-development outcomes might affect these diversification benefits, we run simulations on cohorts of drugs that started development in the same year, between 1998 and 2013.[7] We find that average returns on single-drug portfolios decreased significantly over time, from 2.24 approvals per billion spent in 1998 to 0.46 in 2013. However, portfolios of 20 drugs experienced a much smaller decline in average returns, from 2.55 in 1998, to 1.00 in 2013. This highlights the fact that the benefits of scale for improving capital efficiency have increased over time: in 2013, moving from a scale of 1 to 20 drugs, provides an increase of 115% in average approvals per billion spent (from 0.46 to 1.00), while in 1998 this would have provided only a 14% improvement (from 2.24 to 2.55). Finally, we observe no improvements in portfolio average

[6] Clinical trials costs are drawn randomly from log-normal distributions with parameters calibrated based on cost estimates from DiMasi, Grabowski, and Hansen (2016).

[7] We end the simulations in 2013 to avoid a downward bias in approval frequencies caused by the fact that we can only include completed drug-development programs in our tests.

returns from increasing portfolio size beyond 20 drugs, in any of the years in our sample.

We find a similar pattern in the evolution of portfolio standard deviations: they decrease significantly over time, but substantially less for portfolios of 20 drugs (from 0.64 in 1998 to 0.41 in 2013) than for portfolios of a single drug (from 3.21 in 1998 to 1.14 in 2013). While the decline in standard deviations is partly caused by the decline in approval probabilities, we show that it is also, to a large extent, attributable to a decline in the variance of the number of approvals in the portfolio. This suggests that the increase in concentration around similar targets, which has created larger *pockets* of highly correlated outcomes, is dominated by the significant increase in the total number of targets available, resulting in lower *average* pairwise correlations over time. The overall effect is that portfolios of 20 drugs continue to provide substantial diversification benefits in every year of our sample, reducing portfolio standard deviations by at least 60% compared to the average single-drug portfolio.

Finally, we show that portfolios that are less diversified across biological targets provide significantly worse risk-return tradeoffs. To do so, we build portfolios by randomly drawing from the set of available *targets* each year (as opposed to randomly drawing from the set of unique drugs) and, once a target is selected, we include all the drugs with that target in the portfolio. We continue to do so until the total number of drugs in the portfolio matches the desired scale. We find that, for every year in our sample, this portfolio construction technique results in significantly lower ratios of average return to standard deviation than the random-drug selection approach we use in the rest of our simulations. This finding highlights the idea that improper diversification across targets can substantially reduce the benefits of scale in drug development.

The remainder of this paper is organized as follows: Section 2 provides an overview of the drug development process, highlighting the stages of clinical trials, regulatory requirements, and the roles of different stakeholders. Section 3 describes the data sources and main variables used in our analysis. In Section 4, we analyze the growing importance of financial constraints and their impact on clinical trial terminations, with particular attention to the role of sponsor type and macroeconomic factors. Section 5 explores the increased concentration in drug development programs and the implications for portfolio diversification. Finally, Section 6 presents our simulation results on the performance of drug portfolios and how it has changed over time.

2 The drug development process

The process of drug approval is a multi-stage and highly regulated pathway overseen by regulatory agencies like the U.S. Food and Drug Administration (FDA) or the European Medicines Agency (EMA). It begins with preclinical testing, where a drug's efficacy and safety are evaluated in laboratory and animal studies. This phase ensures that the drug shows promise before human testing. Once preclinical data supports potential success, the drug's "sponsor" (typically a pharmaceutical company, biotechnology firm, or academic research institution) submits an Investigational New Drug (IND) application to the regulatory agency, seeking permission to start clinical trials in humans. Regulatory agencies play a crucial role in evaluating the IND submission, reviewing safety data, and ensuring that the drug can be tested without undue risk to participants.

After IND approval, the clinical trial process begins, consisting of three phases. In phase 1, the drug is tested on a small group of healthy volunteers or patients to evaluate safety, dosage range, and pharmacokinetics. Phase 2 involves a larger group of patients to assess efficacy, further evaluate safety, and finalize dosing for phase 3. In phase 3, the drug is tested on a much larger population to confirm its effectiveness, monitor side effects, and compare it to standard treatments. During each phase, sponsors must submit detailed reports to the regulatory agencies, who may halt trials if concerns about safety or efficacy arise. Sponsors, who are responsible for financing and organizing these studies, may experience significant financial strain as they advance through the lengthy and expensive clinical trial process.

It is important to note that each clinical trial phase can encompass multiple drugs or explore several indications, particularly in areas like oncology or rare diseases, where therapies may target different subsets of patients. In phase 1, the trial typically focuses on a single drug to assess safety and tolerability, often in a small population of healthy volunteers or patients. However, sponsors sometimes conduct "basket" phase 1 trials, testing a drug across different indications or patient subtypes. This strategy helps identify which patient populations or conditions show the most promising safety and efficacy signals, setting the stage for more targeted investigations in phase 2.

As the drug advances to phase 2, sponsors often design multiple trials based on findings from phase 1. A single phase 1 trial might give rise to several phase 2 trials, each tailored to a specific indication, dosage regimen, or patient subgroup. The goal is to further explore efficacy while refining

the target population. This evolution continues into phase 3, where the focus narrows even more. In phase 3, multiple trials may run concurrently, but each will usually have a defined population and indication, based on data from prior phases. The sponsor refines the drug's positioning for regulatory approval by optimizing both the indication and the patient population most likely to benefit from the therapy, potentially seeking approvals across several related indications depending on trial outcomes.

The success or failure of a trial is largely determined by whether its "endpoints" are met. Trial endpoints are the predefined objectives or outcomes that a clinical trial is designed to measure in order to assess the drug's efficacy, safety, or other relevant factors. Endpoints can be categorized as primary (the most important outcomes, like overall survival or symptom reduction) or secondary (additional effects, such as quality of life improvements). Regulatory agencies, like the FDA or EMA, closely evaluate the outcomes of these endpoints when deciding whether a drug can progress to the next phase of development.[8] Although sponsors have the ability to decide whether to proceed with trials, regulatory agencies retain the authority to halt development if the phase 1 endpoints indicate significant safety concerns or if efficacy is insufficient. Sponsors typically decide whether to move forward based on both endpoint results and strategic considerations, but regulatory oversight ensures that only drugs with acceptable safety and efficacy profiles continue to later phases.

Upon successful completion of phase 3, the sponsor submits a New Drug Application (NDA) or a Biologics License Application (BLA) to the regulatory agency. The agency reviews the full body of evidence, including clinical trial data, manufacturing details, and labeling information. The approval process is rigorous, often requiring months of analysis, advisory committee reviews, and potential additional studies before a decision is made. If the drug is deemed safe and effective, it is approved for marketing.

Regulatory agencies continue to monitor the drug post-approval through pharmacovigilance systems, including phase 4 trials, which collect long-term safety and efficacy data in larger and more diverse populations. These post-marketing studies help identify rare or delayed side effects that may not have emerged during earlier trials. In addition to phase 4 trials, pharmacovigilance systems include spontaneous reporting databases (such as the FDA's Adverse Event Reporting

[8] Regulatory agencies also provide input on endpoint selection, especially in phase 3, where they can reject a drug if the endpoint used is not up to standard.

System), electronic health records, and observational studies, all of which help ensure that emerging safety concerns are addressed in real-world clinical settings.

After a drug is approved by regulatory agencies like the FDA or EMA, the firm typically gains certain protections against direct competition. These protections generally fall into two categories: marketing exclusivity and patent protection. Marketing exclusivity is granted by regulatory agencies and prevents other companies from marketing a competing generic or biosimilar version of the drug for a specific period. In the U.S., for example, new chemical entities (NCEs) are granted five years of exclusivity, while biologics receive 12 years under the Biologics Price Competition and Innovation Act. This protection is independent of any patent protection and serves as a reward for the investment in bringing a new drug to market.

Patent protection, on the other hand, is managed through intellectual property laws and can last up to 20 years from the date of filing. Patents cover the drug's chemical structure, manufacturing processes, or methods of use. Because patents are often filed during early development, much of the patent life can expire before the drug is approved. However, firms can sometimes extend patent protection through mechanisms like patent term extensions (for delays in the approval process) or by filing additional patents on new formulations or uses of the drug. While marketing exclusivity offers a shorter period of protection, it begins immediately after approval and blocks generic competition, even if the drug's patent has expired. Both forms of protection provide the firm with a temporary monopoly, allowing them to recoup research and development costs before facing generic competition.

3 Data

We compile data on drug development programs from two proprietary databases: Trialtrove, which contains information on individual clinical trials, and Pharmaprojects which provides data on individual drugs. Both databases are offered by Citeline, which has recently been acquired by Nostrella. The raw data comes in JSON format (downloaded on June 2024 through API calls to the Citeline servers) and contains both numeric fields and text fields which we further process to build structured datasets for our analysis. The remainder of this section describes these datasets in more detail.

3.1 Clinical trial data

Trialtrove is a comprehensive clinical trial database that tracks clinical studies across all phases of drug development globally. Trialtrove compiles its data from multiple sources, including trial registries (such as ClinicalTrials.gov and the EU Clinical Trials Register), regulatory filings, conference presentations, and company reports. While information from trial registries is freely available, the additional sources of information used by Trialtrove play an important role, as they help mitigate the potential sample selection bias created by the fact that sponsors were not required to report information on their clinical studies until recently (e.g. 2007 in the United States, and 2022 in the European Union).

Trialtrove tracks over 170 pieces of information for each individual clinical trial, including data on the drug and disease targeted in the trial, as well as the trial's beginning and end dates, phase, sponsors, locations, principal investigators, patient segments, endpoints (objectives) and outcomes. The trial beginning and end dates allows us to analyze trends in clinical research over the past three decades. The information on trial outcomes (whether the trial endpoints were met or not, or whether the trial was terminated and, if so, for what reason) enable us to analyze the extent to which trials are terminated due to lack of funding or other business-related reasons (Section 4). The database does *not* contain information on whether the trial lead to a successful transition to a subsequent phase (or approval). This is inferred from the drug-level data described in the following two sections.

Table 1 presents summary statistics for some key variables in the Trialtrove database. Panel A show coverage statistics for the full dataset (first two columns), as well as for the top 40 countries where the most trials were conducted (the following two columns), and for the United States (last two columns). The full sample of 200 countries consists of 407,935 unique trials, with 374,191 of them (92%) coming from the top 40 countries, and 123,656 of them taking place in the United States (30%). Panel A also shows that, while some clinical trials are missing critical information such as the trial phase, sponsor, and therapeutic area (e.g. disease targeted), coverage for these items is quite good, ranging from 88% to 93% in the full sample. Panel B shows that around 14% of the trials in our dateset are still ongoing, 74% have completed and 11% have been terminated, though, in the United States, a larger proportion of trials seem to be terminated (17%). In our

analysis below, when data on trial outcomes is needed, we discard the ongoing trials since their outcomes are not yet known.

The remaining panels in Table 1 provide more detailed information on trial phase, sponsor type, and therapeutic area. Panel A shows the breakdown of the data by trial phase. Phase 3 is the least common, reflecting the fact that most drugs never reach that phase. Note that it is not surprising that we have more phase 2 trials than phase 1 trials since, as discussed in the previous section, a successful phase 1 trial often results in multiple phase 2 trials often looking for signals of efficacy in a number of possible settings (e.g. different disease areas). While we have a substantial number of phase 4 trials in our data, we do not use them in our subsequent analysis, since these are post-approval trials, and our study is primarily focused on the costs and risk of obtaining regulatory approval for drug-development projects. Panel D shows the breakdown by sponsor type (as categorized by Trialtrove). The main takeaway from this panel is that, while large "Top 20 Pharma" companies sponsor a substantial proportion of trials (20%), the vast majority of trials are sponsored by academic institutions (56%) and smaller pharma companies (23%). Not surprisingly, in the United States, where the financial incentives of obtaining approval are stronger, the top 20 pharma companies sponsor a larger proportion of trials (34%). Finally, Panel E breaks down the sample by the broad therapeutic area of the disease targeted in the trial. Here we see that Oncology and Central Nervous System (CNS) diseases are the most widely represented, together accounting for 44% of the trials in our data (55% in the United States), though Autoimmune/Inflammation, Infectious Disease, Cardiovascular and Metabolic/Endocrinology are also very well represented at 13-14% of the sample each.[9]

Figure 1 shows trial counts over time, by phase. On the x-axis, we have the year when the trial started. We include only the data in the past 3 decades to keep the graph more readable, though we do have a small number of trials starting before 1993. The top panel includes all the countries in our dataset, and the bottom panel includes only the trials taking place in the United States. In both panels, we observe a steady increase in the number of clinical trials (for all phases) over time, though, in the United States this growth seems to have stopped around the year 2007 when the FDA mandated that clinical trial data be made public. The graph shows a noticeable increase in

[9]Note that, for panels D and E, these percentages will add up to more than 100% as a trial can have multiple sponsors or target multiple therapeutic areas.

clinical trials around the onset of the COVID-19 pandemic. The decrease in trials following this period is likely caused by the fact that Trialtrove may still be collecting data on the trials that started in 2022 and 2023.

In Table 2, we highlight a key feature of drug-development projects, namely their lengthy durations. The table presents average trial duration, in years, by phase and therapeutic area, for all countries in our sample (Panel A) as well as just for the trials taking place in the United States (Panel B). The last column in the table shows the total duration of phases 1 through 3 combined. This column shows that, averaging over all therapeutic areas, the approval process takes 6.05 years (7.18 years in the United States). There is substantial variation on trial duration across therapeutic areas, ranging from an average of 3.99 years for Vaccines to 10.12 years for Oncology drugs. Phases 2 and 3 have similar durations, at 2.49 and 2.37 years respectively (2.76 and 2.55 in the United States), and they take substantially longer to complete than phase 1 trials (1.19 years in the full dataset, and 1.89 years in the United States).[10] The results in Table 2 attest to the fact that drug-developing companies experience very substantial delays between the time costs are incurred and the time revenue (if any) is realized.

3.2 Drug-level data

The Pharmaprojects database tracks drug development pipelines globally, offering extensive coverage of drugs from preclinical research through to post-marketing phases. It aggregates data from a variety of primary sources, including regulatory filings, corporate disclosures, press releases, and scientific literature. Pharmaprojects provides detailed profiles on each drug, tracking close to 100 pieces of information including the drug's chemical structure, mechanism of action, therapeutic target, and commercial status. Each drug profile also contains information on all the diseases for which the drug has been tested, and on the important parties involved in its development (e.g. its originator and licensees).

Table 3 shows summary statistics for the drug-level data obtained from Pharmaprojects, both for the full sample of drugs (first two columns) and for the sample of drugs for which we have clinical trial data from Trialtrove (last two columns). Panel A shows that we have 104,129 unique

[10] To account for the presence of outliers, in Table A.1 of the Appendix, we present median trial durations by therapeutic areas. We find that the median drug approval process takes 4.3 years in the full sample, and 5.37 years in the United States.

drugs in Pharmaprojects, out of which 28,606 also have clinical trial information. Panel A also shows that almost all drugs in the dataset also have information on the drug's origin (e.g. chemical or biological) and the broad indication (disease) groups for which the drug has been developed. Only 70% of drug profiles contain information on the drug's mechanism of action (MOA) and 46% of them contain information on the drug's biological target. These variables a more readily available in the sample of drugs with clinical trial data, at 83% and 63% respectively. We discuss drug MOAs and targets in more detail in Section 5 where we analyze the industry's concentration around similar MOA's and drug targets.

Panel B of Table 3 shows that about a third of the drugs in Pharmaprojects have biological origin (aka "biologics" or "large molecule" drugs) and two thirds have chemical origin (aka "synthetic" or "small molecule" drugs). Panel C shows a breakdown of drugs by the main therapeutic area (indication group) of the diseases they target. Just as with the clinical trial data, we see that drugs targeting Anticancer and Neurological disease appear with the highest frequency in the drug-level data as well. Comparing this panel with Panel E in Table 1, also shows that Pharmaprojects and Trialtrove use different disease classifications (and names). This complicates the process of merging the two databases at the drug-disease level. We discuss this merging process in more detail in Section 6.

Most of the Pharmaprojects data is not timestamped. However, for most drugs that have ever been approved (for any disease, in any country), the database provides a "marketing" text field, which contains a discussion (by Pharmaprojects analysts) of when the drug was approved for each disease, in each country. The database also includes a "keyEvents" field, which provides timestamps for some (but not all) important events in the drug's global development (e.g. approval dates in particular countries, or dates when new licensing agreements for the drug were created). We use a combination of natural language processing algorithms and manual verification to extract these approval dates from the "marketing" and "keyEvents" sections.

3.3 Drug-development programs

In 2020, Pharmaprojects performed a large scale, manual curation exercise through which they were able to identify which sponsors developed which drug, for which disease, in which country. Pharmaprojects refers to this as the "Drug Program Landscape" dataset, and to each drug-disease-

sponsor-country record as a "drug program". We adopt this definition in our study. For the purpose of our analysis, the most important variable in this dataset is the "highestPhaseReached" variable, which, as the name suggests, records the highest development phase reached by each drug-development program. This is the main variable we use to determine which clinical trial phases a drug has passed or failed. Namely, we assume that every clinical trial associated with a given drug-development program was successful if its phase is lower than the "highestPhaseReached" of that program, and unsuccessful otherwise.

Table 4 presents descriptive statistics for the Drug Program Landscape dataset described above. Panel A shows that this dataset contains information on 584,726 unique drug–disease–sponsor–country records, corresponding to 226,356 unique drug–disease–sponsor triples, and 174,937 drug-disease pairs. Panel B breaks down the unique drug–disease–sponsor–country records by their current development status (as of June 2024, when we downloaded the data) and shows that 29% of them are "Widely Launched" (i.e. they have been approved and is already selling), 55% of them have "Ceased" (i.e. they have ended development and have either failed or are awaiting approval) and 17% of them are still under active development. Note that the 29% approval frequency documented in this panel is not representative of the ex-ante probability of a drug being approved for a given disease, since drug-disease pairs that receive approval in some country, are significantly more likely to be subsequently developed in other countries and/or by other sponsors (e.g. through licensing agreements). In the last two columns in the table we compute these frequencies only using drug-development programs from the United States. Here we see that only 6% of these programs have launched, which, as we will see below, is much more representative of the average probability of success in the United States.

Panel C of Table 4 shows a breakdown of drug-development programs by "highestPhaseReached". Here, under the "Approved" category, we combine both drugs that have reached the "Registered" phase (i.e. approved but not yet selling) and the drugs that have reached the "Launched" phase (approved and selling). Once again, we see a much higher frequency of approved drugs in the full dataset (32%) compared to the United States (7%), again stemming from the tendency of successful drugs to be developed in more countries than unsuccesful ones.

In Table 5, we use the "highestPhaseReached" to estimate average transition probabilities for each phase of the approval process, by therapeutic area. We do so using both the entire dataset

(Panel A) as well as just the United States (Panel B). In Panel A, to avoid the double-counting issues mentioned above, we aggregate the data across countries and sponsors, by taking, for each drug-disease record, the maximum of "highestPhaseReached" over all the countries and sponsors associated with that drug-disease pair.[11] This raises the concern that low regulatory standards in some of the countries in our dataset could artificially inflate our transition probabilities (e.g. if a drug-disease program fails in the United States and the European Union, but gets approved in any other country, we would count that as a successful program). To avoid this concern (beyond providing results for the United States alone), in Panel A, we use only the countries with the very highest regulatory standards: the United States, Japan, and the European Union countries.[12] The first column in the table shows the number of unique drug-disease records for each therapeutic area as well as in the full sample (last row in each panel). The following 4 rows in the table show the frequency with which drug-disease programs pass each phase, from preclinical ("Pre" in the table) all the way up to approval ("A" in the table). More specifically, the probability to transition from phase i is computed as 1 minus the probability of failing to transition from phase i (i.e. 1 minus the number of drug-disease pairs that fail in phase i divided by the number of drug-disease pairs that reach at least phase i):

$$\mathbb{P}_i = 1 - \frac{\text{N(highestPhaseReached = i)}}{\text{N(highestPhaseReached} \geq \text{i)}} \tag{1}$$

where $i \in \{Pre, 1, 2, 3, A\}$.[13]

The results in the "Total" row from Panel A of Table 5 show that a substantial percentage of drug development programs (60%) never pass the preclinical stage. During the clinical trial process, most failures happen in phase 2 (only 46.8% of drugs pass it), followed by phase 3 (63% transition probability) and phase 1 (70% transition probability). The last two columns in the table present overall probabilities of success (PoS) from preclinical to approval ($\mathbb{P}_{Pre-A}$, which is the product of all 4 phase transition probabilities) and from phase 1 to approval ($\mathbb{P}_{1-A}$, which is the product of $\mathbb{P}_{1-2}$, $\mathbb{P}_{2-3}$, and $\mathbb{P}_{3-A}$). These estimates show that, the unconditional PoS starting from the

[11] Given this aggregation, in the rest of the paper, a drug-development program is synonymous with a drug-disease pair.

[12] We verify, in robustness tests, that all our results are qualitatively unchanged if we use the full set of countries in the Pharmaprojects dataset to determine likelihoods of approval.

[13] In calculating these probabilities, we exclude drugs that are still in active development.

preclinical phase is only 8.3% and the PoS starting from phase 1 is 20.8%. For the United States (Panel B), the transition probabilities and PoS estimates are lower almost across the board, though the largest difference is in phase 3, where the average transition probability is only 56.5% in the United States, compared to 63% worldwide.

Table 5 also shows substantial variation in PoS across different therapeutic areas. For example, anticancer and neurological drugs, which together constitute almost half the sample, have the lowest PoS (8.4% and 21% respectively for the $\mathbb{P}_{1-A}$). On the other end of the spectrum, anti-infective drugs, also very well represented in our sample, have a PoS starting from phase 1 of 32.2%. Overall, the results in Table 5 highlight another key characteristics of the drug-development process, namely that the likelihood of the project ever generating any revenue is very low.

4 The growing importance of financial constraints

The long durations and low probabilities of success of drug-development projects, together with the fact that a large proportion of these projects are undertaken by sponsors that do not have a significant source of internally generated cash flows, suggests that innovation in drug development is significantly exposed to shocks in the supply of external financing. In this section, we present evidence that this exposure has likely increased over time. We show that, over the past two decades, the percentage of trials terminated due to lack of financing has substantially increased, the percentage of trials with large, "Top 20 Pharma" sponsors has decreased, and clinical trials exhibit significantly longer durations, and higher scientific risk. We then show regression evidence tying funding-related trial terminations to sponsor size, trial durations, equity valuations and macroeconomic uncertainty.

We begin with an analysis of trial outcomes, as reported in the "trialOutcomes" field in Trialtrove. Table 6 summarizes how Trialtrove analysts have categorized the outcomes of completed (Panel A) or terminated trials (Panel B). For each outcome category, the first two columns report counts and frequencies for that outcome in the full sample, and the last two columns present these statistics for trials taking place in the United States. Panel A shows that, for a significant portion of trials, outcome information was either not found (33% of trials have "Outcome unknown") or was difficult to categorize (11% of trials have "Outcome indeterminate"). This caveat notwith-

standing, Panel A suggests that almost half of completed trials (45% in the full sample, and 48% in the United States) have positive outcomes (i.e. primary endpoints were met). In contrast, only 9% of completed trials have negative outcomes in the full sample, and 12% in the United States. While it is possible that this discrepancy between positive and negative outcomes could be caused by information on trials with negative outcomes being more difficult to find, it is also important to remember that, meeting primary endpoints does not necessarily imply a successful transition to a subsequent phase, since, as described in Section 2, trial results still need to be submitted to regulatory agencies for evaluation.[14]

More importantly for our analysis, Panel B of Table 6 shows that, despite a large number of trials being terminated for reasons having to do with the science and logistics of the trial (e.g. "Poor enrollment", "Lack of efficacy", and "Safety/adverse effects"), a substantial number of trials have also been terminated for financial or business-related reasons. Indeed, the first column in the table shows that 1,268 trials have been terminated due to lack of funding, and over 8,300 trials have been terminated based on business decisions (3,220 terminations due to "Pipeline reprioritization", 1,237 due to "Drug strategy shift" and 3,917 terminations citing "Other" business decisions).[15] The last two columns in Panel B show that these financing or business-related terminations are relatively more common in the United States than in the full sample, which is perhaps not surprising, since the larger healthcare costs in the United States are likely to result in substantially larger costs of running clinical trials (Lorenzoni and Dougherty (2022)).

To examine the extent to which these funding- and business-related terminations have changed over time, in Figure 2, in the two leftmost panels, we plot the percentage of trial terminations due to lack of funding and, in the rightmost panels, we plot the percentage of terminations citing a business decision ("Pipeline reprioritization", "Drug strategy shift", or "Other"). The top-left panel shows a strong upward trend in the percentage of funding-related terminations both in the full sample, and particularly in the United States, where, in 2023, 7.8% of trial terminations cited lack of funding, up from only 2.2% in 2000. Consistent with the idea that these terminations are caused by financial constraints, the bottom-left panel shows that funding-related terminations (in

[14] Also note that, even if trials meet their endpoints, sponsors might decide not to continue development if they believe trial results were not sufficiently positive to warrant a subsequent phase.

[15] The difference between "Pipeline reprioritization" and "Drug strategy shift" according to Pharmaprojects is that with a pipeline reprioritization, the compound is no longer in active development for any indication, while with a drug strategy shift, development continues in other indications or with a follow-on compound.

the full sample) are significantly less frequent among Top 20 Pharma sponsors than in the "All but Top 20 Pharma" group, though both groups show a significant upward trend. This is direct evidence that financial constraints play an increasingly important role in the drug development process.

The top-right panel of Figure 2 also shows an increase in the percentage of terminations citing business reasons, though most of this increase seems to have happened after 2020. It is conceivable that some of these terminations could also be funding related (perhaps a "Pipeline reprioritization" would not be needed if funding was readily available). While this is certainly possible, the bottom-right panel of the figure shows that financial constraints are likely not the main driver behind these business-related terminations, given that they are significantly more common among the "Top 20 Pharma" sponsors.

Next, we examine the extent to which the composition of trial sponsor types has changed over time. To this end, in Figure 3 we plot the number of trials initiated every year, by sponsor type. We restrict ourselves to the largest 3 sponsor categories, which cover between 88% and 99% of trials in our sample every year: "Academic", "Top 20 Pharma", and "All Other Pharma". The primary insight obtained from this figure is that the "Top 20 Pharma" sponsors have significantly decreased their involvement in the clinical trial process over time, while the "All Other Pharma" sponsors have substantially increased theirs. More specifically, Panel A shows that, in 2023, the largest pharma companies sponsored only 11% of all trials in our sample, down from 28% in 2003. In contrast, pharma companies outside the top 20 sponsored 31% of trials in 2023, up from 13% in 2003. Panel B shows that these trends are also present when we include only trials from the United States. The percentage of trials sponsored by "Top 20 Pharma" nearly halved in the last 20 years (from 43% to 24%) while the percentage of trials sponsored by "All Other Pharma" nearly doubled (from 18% to 35%). Finally, while relatively stable over time, the percentage of trials with academic sponsors is substantial, at 42% in the full sample, and 36% in the United States in 2023. These findings highlight a key structural change in the drug-development process: the past two decades have seen a steady and significant increase in the proportion of trials sponsored by entities which are more likely to experience financial constraints.

To explore this phenomenon further, in Figure 4 we break down the trends documented in Figure 3 by trial phase. The figure presents three main findings. First, the general decrease (increase)

in trial sponsorship by large (small) pharma companies is present at all phases of development. Second, these trends seem to be the somewhat stronger for phase 1 trials, and weaker for phase 3 trials. Indeed, by 2023, "Top 20 Pharma" sponsorship has decreased to only 10% for phase 1 trials while "All Other Pharma" sponsorship has increased to 40% of trials. For phase 3 trials, these percentages are somewhat less extreme at 17% and 32% respectively. Third, trials in the United States see a significantly larger involvement from "Top 20 Pharma" companies in all three phases, and particularly in phase 3, where they still sponsor 37% of trials. This is not surprising, given the larger financial rewards that come with approval in the United States (Mulcahy, Schwam, and Lovejoy (2024)).

To investigate how trial durations have changed over time, from 2001 to 2023, we calculate average trial durations by trial phase and therapeutic area, separately for each cohort of trials ending in each year.[16] For each therapeutic area, every year, we add up the average trial durations for phases 1 through 3, and we plot these total durations over time in separate panels of Figure 5. For almost all therapeutic areas, we observe a significant increase in trial durations over time, particularly for trials in the United States. For example, over the past 20 years, in the United States, the average duration from phase 1 through phase 3 has increased by over 30% for Oncology drugs, and has almost doubled for CNS, Autoimmune/Inflammation, Cardiovascular, Metabolic/Endocrinology, and Genitourinary drugs. While Infectious Disease drugs have seen a reduction in trial durations around the COVID-19 pandemic, they seem to have resumed their upward trend since. Finally, while Vaccines and Ophthalmology drugs do not show as strong an increase in trial durations over time, it is important to remember that they only account for about 6% of the trials in our sample (see Panel E in Table 1).[17]

Next, we investigate the extent to which the scientific risks associated with the approval process have changed over time. While such risks are certainly difficult to measure, we believe it is reasonable to assume that trials without a previously approved target or mechanism of action (MOA) should exhibit substantially higher uncertainty with respect to their likelihood of approval. As such, in the top three panels of Figure 6, we plot the percentage of trials investigating drugs with targets or MOA's with no prior approval. In the bottom three panels, we express these fre-

[16] We start in 2001 to ensure enough observations of trials *ending* in that year, for each therapeutic area and trial phase.

[17] Figure A.1 in the Appendix shows very similar trends in median trial durations as well.

quencies as a percentage of the number of unique drugs investigated in trials starting each year. The two leftmost panels show almost a doubling in the preponderance of trials and drugs with no previously approved targets and MOAs. Comparing the middle 2 panels with the two rightmost panels we see that this trend is significantly more pronounced in the trials sponsored by the "All but Top 20 Pharma" group. The results in Figure 6 suggest that the scientific risk involved in the drug-development process has significantly increased over the past two decades, particularly for drug-development programs undertaken by smaller sponsors.

The trends documented in this section provide suggestive evidence that innovation in drug-development has become increasingly more susceptible to underinvestment risk. The larger preponderance of small sponsors (Figure 3 and 4), dramatically longer lags between cash outflows and inflows (Figure 5), and increased prevalence of unproven drug targets and MOA's (Figure 6) are certainly consistent with the upward trend in funding-related trial-terminations documented in Figure 2. We conclude this section with a regression analysis investigating the extent to which funding-related terminations are related to fluctuations in equity valuations and macroeconomic uncertainty controlling for the common trends documented above.

In Panel A of Table 7, we present trial-level regressions where the dependent variable is binary indicator which equals 1 if the trial was terminated due to lack of funding (as specified in the "trialOutcomes" field in the Trialtrove dataset). The regression includes all trials that were terminated between 2000 and 2023. The independent variables include a linear trend, an indicator for whether the trial had a "Top 20 Pharma" sponsor, the trial's duration, the Baker and Wurgler sentiment index as a measure of supply shocks in the equity market, and the CBOE's VIX index as a measure of uncertainty shocks in the equity market.[18] The results in Panel A show that, consistent with the findings documented in this section, funding-related terminations exhibit a strong upward trend, are significantly less likely if the trial has a "Top 20 Pharma" sponsor, and significantly more likely for trials of longer duration. The last two columns in Panel A show that, even after controlling for the trend, sponsor type, and trial duration, terminations due to funding are significantly less likely during periods of high sentiment in equity markets, and more likely during periods of high expected volatility in equity markets.

In Panel B, we present macro-level regressions where the dependent variable is the percentage

[18]The daily VIX index was averaged out over the 12 months leading up to each trial's termination.

of trials terminated due to lack of funding in each year between 2000 and 2023. Since the trends in trial durations and sponsorship type are highly correlated with our linear trend, we exclude them from these macro-level regressions, and only control for the linear trend. Given the propensity of small biotech firms to be financed through equity initial public offerings (IPO's), we also include the aggregate IPO volume in the United States as a control variable. The results in Panel B, not surprisingly also show a strong linear trend in the percentage of terminations caused by lack of funding. Controlling for this trend however, we find that the percentage of funding-related terminations is significantly lower when equity-market sentiment is high, when aggregate IPO volume is high and when expected equity-market volatility is low. The last column in Panel B shows that the Baker-Wurgler index remains significant even when controlling for IPO volume, despite the fact that it uses IPO volume as part of its construction. This suggests that current IPO volume is not a sufficient statistic describing the impact of equity supply shocks on trial-termination decisions.

5 The increased concentration in drug-development programs

The previous two sections have laid out the case that the financing frictions faced by firms in the drug-development process are substantial, and have become significantly more severe over time. The underinvestment risk caused by these frictions is particularly concerning given that successful drugs stand to have significant societal value beyond their potential profitability for firms. The fact that this underinvestment risk seems to have increased at the same time that artificial intelligence algorithms have begun to significantly expedite the process of uncovering promising new drug candidates (Sarkar et al. (2023), Chen et al. (2024)) makes it even more imperative that we explore alternative approaches of alleviating the financial constraints faced by drug developers.

We focus our attention on the securitization approach proposed by Fernandez, Stein & Lo (2012), who argue that a "megafund" could be used to finance a large-enough number of drug-development programs that the scientific risks associated with individual projects would largely be diversified away. This could make the risk-return profile of the fund attractive enough to raise more capital, at a better price than if drugs were funded separately. The fund could also issue different tranches of debt and equity, allowing for the distribution of different levels of risk to market

participants of different levels of risk aversion. However, the 2007-2008 financial crisis has made it abundantly clear that securitization without a thorough understanding of the correlations between constituent assets can have severe consequences. In this section, we take the first steps towards understanding correlations between drug-development programs. We begin with a discussion of how these correlations might arise, and we continue with an analysis of how they have evolved over time.

Drug discovery entails developing molecules or techniques designed to disrupt mechanisms of disease and/or repair biological processes in disarray. Research initially focuses on developing an understanding of the pathological basis of the disease, and continues with a search for approaches to modify biologic pathways through which the disease progresses. These pathways – targets – are validated through the successful development of a safe and effective therapeutic. Failure of drug development comes from either the selection of a target that is non-fundamental to the disease state, a drug that does not impact its target to a meaningful extent, or a product that is overly toxic. Common exposures to these failure modes can result in substantial correlations between the probability of success (PoS) of drugs with similar target selection, mechanism of action or molecular structure. As concrete examples of such correlations, consider that 16 Insulin-like Growth Factor Receptor-1 inhibitors have been studied in 183 separate clinical trials, and all of these trials have failed, indicating a flawed target hypothesis. On the positive side, at least nine PD-1/L1 monoclonal antibodies have received FDA approval, following initial trials for these classes of molecules successfully validating the target, which led many developers to attempt to create molecules that are very similar to the initial successful drugs.

This inherent correlation between drug development programs seems to be readily acknowledged in the drug-discovery process. This is evidenced by the fact that formal computational methods have been developed specifically for the purposes of identifying drugs that are suitable for therapeutic substitution and drug repurposing, presumably to take advantage of this implied correlation. However, very little work has been done to assess the effects of this correlation on drug valuation and financing.[19] We attempt to at least partially fill this void in the literature, by investigating

[19] A notable exception is Lo and Siah (2021) who show that the performance of portfolios of rare diseases decreases with the pairwise correlation in transition probabilities between drugs. The authors note that they were not able to find any studies that try to quantify this correlation and assume it is 0.2 for any two drugs. Our study aims to partially fill this gap in the literature by explicitly investigating the extent to which the average pairwise correlation in approval outcomes has changed over time.

the extent to which drug-development programs cluster around similar biological targets and how this clustering has evolved over time.

We begin, in Figure 7, with an illustration of drug-target concentration in the full Pharmaprojects sample. Each node in the figure is a different biological target found in the database, and the size of each node is proportional to the number of drugs with that target in the sample.[20] There are 47,546 unique drugs with target data in the sample, and 3,464 unique targets, for an average of 13.7 drugs per target. To quantify the degree of concentration of drugs around the same target, we compute the Gini index of inequality in node sizes, where a Gini index of 0 would signify that each target is used by an equal number of drugs, and an index of 1 would mean that all drugs have the same target. We find an Gini index of 0.78 in the full sample, which suggests there is a high degree of concentration in target selection in the data.

To analyze how this target concentration has evolved over time, we merge the drug-level data (Pharmaprojects) with the trial-level data (Trialtrove) based on "drugId" and we keep only the phase 1 trials, to ensure that we are not double counting drugs that are more successful (i.e. drugs that have passed more phases in the clinical trial process). We then repeat the process used to create Figure 7 using only drugs that started phase 1 trials in 1998 (the first year in the sample with at least 100 unique targets) and then using only drugs that started phase 1 trials in 2023. We plot these graphs in the top two panels of Figure 8. These panels show a significant increase in the number unique targets (nodes) in the past 25 years, but also increased concentration of drugs around the same target (larger inequality in node sizes in 2023 compared to 1998). The bottom two panels of Figure 8 illustrate trial-target concentration (i.e. node sizes are proportional to the number of trials investigating drugs with that target). These bottom panels show a very similar increase in concentration of trials around the same target.[21]

Figure 9 presents a more detailed look at the evolution of drug- and trial-target concentration over time. The top-left panel shows the number of unique trials starting phase 1 each year, as well as the number of unique drugs being investigated in those trials, and the number of unique targets of those drugs. This panel shows that the increase in the number of trials and drugs has

[20] The positioning of each node in space is not meaningful.

[21] The Gini coefficients in the bottom panels of Figure 8 are lower than the Gini coefficient in Figure 7 because they are based only on drugs starting phase 1 in each year, while Figure 7 uses all the drugs in our entire sample (across all years and all phases).

outpaced the increase in the number of targets. The top-right panel makes this observation more precise, as it shows that the number of trials initiated per target has increased from 1.3 in 1998 to 2.4 in 2023. Similarly, the number of drugs investigated per target has nearly doubled, from 1.04 in 1998, to 1.99 in 2023. In the bottom left panel, we separately compute the Gini index of inequality in drug-target selection (as discussed in Figure 7) using only the drugs starting phase 1 trials each year. The plot shows an 85% increase in drug-target concentration (from 0.21 to 0.39) in the past 25 years. The bottom right panel computes these Gini coefficients at the trial-target level and shows a similar trend in concentration (a 36% increase, from 0.425 in 1998 to 0.578 in 2023).

Next, we investigate the extent to which drugs with common targets have correlated outcomes in the clinical trial process. To this end, we use the Pharmaprojects data on drug-development projects described in Section 3.3 and keep all projects that have reached at least phase 1, for which we have drug target data. We then calculate, for each drug i, the average approval rate (AAR) of all other drugs using the same target:

$$AAR_i = \frac{1}{|S_i| - 1} \sum_{j \in S_i, j \neq i} Approved_j \tag{2}$$

where S_i is the set of drugs sharing the same target as drug i, and $Approved_j$ is an indicator that drug j was approved. The term $|S_i|$ represents the number of drugs in the set S_i. To ensure that $|S_i| > 1$ for each drug, we restrict the sample to drugs with targets developed by at least two drugs.

Table 8 presents results from regressions of the type:

$$Approved_i = \alpha + \beta AAR_i + \gamma X_i + \epsilon_i \tag{3}$$

where X_i contains various drug-level controls which we will discuss shortly. We estimate the model using OLS for simplicity, though the results are qualitatively unchanged when we use logistic regressions. Standard errors are clustered at the therapeutic-area level (i.e. indication groups as illustrated in Table 3). In the first column of Table 8 we use the AAR_i as the single independent variable and we find that it a significant predictor of approval outcomes. Specifically, a one

percentage point increase in the approval rate of drugs with the same target is associated with a 0.94 percentage point increase in the likelihood of the drug being approved. Furthermore, the R^2 coefficient shows that AAR_i by itself explains 20.5% of the variation in approval outcomes.

In columns 2 through 4 of Table 8 we add several other drug-level characteristics to our regression. Specifically, in column 2, we control for the therapeutic area of the drug (through indicator variables). In column 3 we control for the target family of the drug by including dummy variables which equal 1 if the drug's target is an "enzyme", "receptor", "transporter", "ion channel", or "cytokine/growth factor" (as defined by the "targetFamilies" variable in Pharmaprojects). Similarly, in column 3, we control for the general MOA family of the drug as defined by the "mechanismHierarchy" variable in Pharmaprojects. Specifically, we include indicator variables for each of the following MOA categories: "immunosuppressant", "immunostimulant", "dna inhibitor", "dna synthesis inhibitor", "ion channel antagonist", "protein kinase inhibitor", "growth factor receptor antagonist", "cell cycle inhibitor", "apoptosis stimulant", "angiogenesis inhibitor", and "5 hydroxytryptamine receptor antagonist". The results in columns 2 through 4 show that the AAR_i variables remains statistically significant throughout, which provides reassurance that it is not simply capturing general variation in approval rates by therapeutic area, target family or MOA family.

The results in this section highlight two competing forces that may act to influence the degree to which portfolios of drug-development programs can result in better risk-return tradeoffs. On the one hand, the increased concentration of drugs and trials around similar targets implies that randomly selecting drugs for the portfolio will result in a higher likelihood of selecting drugs with the same target, which, based on the findings in Table 8, should result in more correlated outcomes across drugs in the portfolio. On the other hand, the larger number of targets available could improve portfolio diversification through careful selection of drugs with unrelated targets. We examine the effect of these competing forces on drug-portolio performance in the next section.

6 Performance of drug portfolios

We begin our analysis of drug-portfolio performance with a series of simulations aimed at developing a better understanding of how portfolio returns and volatilities vary with the number of drugs in the portfolio. In particular, we are interested in a measure of portfolio returns that captures societal

(rather than financial) value as well as possible. As such, we measure portfolio returns as the number of drugs in the portfolio that achieve regulatory approval, divided by the total clinical trial costs incurred by the drugs in the portfolio. While we acknowledge that not all approved drugs provide equal societal value, a simple approval count avoids the need to pass judgement on the relative improvements in quality of life provided by the alleviation of different medical conditions. Roughly speaking, one could also think of these as the total financial returns obtained if, on average, approved drugs resulted in $1 billion of free cash flows (in present value terms).

6.1 Full-sample simulations

For our initial set of simulations, we use the Pharmprojects drug-development project dataset described in Section 3.3, and we make the simplifying assumption that all these projects would have been available to invest in at the same time (we relax this assumption later on in this section). For each N from 1 to 100, we randomly draw N drugs from the entire dataset, and use the "highestPhaseReached" variable to record the outcomes of the portfolio's drugs. We repeat this 100,000 times for each N. For each simulated portfolio of N drugs, we compute the return on that portfolio as:

$$R_N = \frac{\sum_{i=1,N} Approved_i}{\sum_{i=1,N} Cost_i} \tag{4}$$

where $i = 1, N$ are the N drugs in the portfolio, $Approved_i$ is an indicator that drug i was approved, and $Cost_i$ is the total clinical trial costs incurred by that drug (in billions of 2023 dollars). Since our dataset does not provide data on these costs, we draw them randomly as well, using the estimates from DiMasi, Grabowski, and Hansen (2016) to calibrate our simulations. The authors surveyed pharmaceutical companies to obtain detailed research and development costs for a set of 106 randomly selected drugs. They document average out-of-pocket development costs of 25.3, 58.6, and 255.4 million USD (in 2013 dollars) for phases 1, 2, and 3 respectively. The standard deviations of these costs were 29.6, 50.8, and 153.3 respectively. We convert these parameters to 2023 dollars, and, each time we draw a random portfolio of N drugs for our simulations, we also draw their development costs from log-normal distributions with means and standard deviations

as listed above.[22]

Figure 10 shows histograms of the simulated number of approved drugs per portfolio (left), total portfolio costs (middle) and the ratio of the two i.e. R_N (right) for portfolios of 1 (top), 20, (center), 40 and 100 drugs (bottom). The top three histograms show that, with portfolios of a single drug, we obtain very skewed distributions for approvals, costs, and returns, with a particularly concerning mass at 0 for the distribution of returns (covering 79% of outcomes). The histogram legends contain information about the mean and standard deviations of the distributions depicted. Portfolios of 1 drug provide an average of 0.21 approvals per portfolio, cost an average of \$0.2 billion to develop, and provide an average of 0.57 approvals per billion spent.

The second row of histograms shows that even moderately sized portfolios of 20 drugs provide significant diversification benefits over portfolios of a single drug. While the average number of approvals and average portfolio costs for $N = 20$ are, naturally, 20 times larger than their $N = 1$ counterparts, the standard deviations are only 4.4 times larger for the number of approvals (from 0.41 for $N = 1$ to 1.81 for $N = 20$) and 4.5 times larger for total costs (from 0.21 for $N = 1$ to 0.95 for $N = 20$). The benefits of diversification are even more apparent when we compare the rightmost histograms: the standard deviation of returns (R_N) decreases by 70%, from 1.24 for $N = 1$ to 0.38 for $N = 20$, and the likelihood of the portfolio obtaining no approvals decreases from 79% when $N = 1$ to less than 1% when $N = 20$. Finally, one of the key findings of our paper, which we will discuss in more detail later on in this section, is that this decrease in risk is associated with an almost doubling in average returns, from 0.57 drugs per billion spent for $N = 1$ to 1.05 for $N = 20$.[23]

The bottom row of histograms in Figure 10 compares portfolios of 40 drugs with portfolios of

[22] Fernandez et al. (2012), the original "megafund" paper that first suggested securitization as a solution to financing drug development, also includes simulations showing that portfolios of drugs are likely to provide attractive risk-reward profiles. A series of follow-on papers further refine these simulations and apply them to specific therapeutic areas (Fagnan et al. (2014)), Fagnan et al. (2015), Das et al. (2018), Chaudhuri et al. (2019), Siah et al. (2021). We differ from these studies in at least three important respects. First, while they draw transition probabilities from synthetic distributions calibrated on previous studies, we draw directly from the distribution of realized outcomes. This allows us to account for the inherent correlation in outcomes in the data. Second, we focus on the number of approvals per billion spent, as opposed to financial returns. This allows us to hone in on the societal value provided by drug-portfolio financing, and also sidesteps the complexities associated with determining the appropriate discount rates that should be applied to the cash flows generated by the portfolio. Finally, as discussed in more detail in the following section, we provide a time-series analysis of the extent to which the risk-return profiles of drug-portfolios have changed over time, by running simulations separately on cohorts of drug-development programs that started clinical trials in the same year.

[23] Note that this is a mean of ratios not a ratio of means (which would not vary with the size of the portfolio).

100 drugs. Not surprisingly, the average number of approvals and total costs increase linearly with portfolio size. However, the bottom right panel shows that the average portfolio return is virtually identical between portfolios of size 100, 40, and 20 (1.05 for all three) suggesting that we obtain virtually no additional benefits in terms of average returns by increasing portfolio size beyond 20 drugs. On the other hand, comparing standard deviations reveals that increasing portfolio size does continue to significantly decrease portfolio risk, for example providing more than 50% reduction in standard deviation from $N = 20$ to $N = 100$. To showcase the benefits of this reduction in portfolio risk with a focus on the left side of the distribution, we note that our simulations estimate that the probability of obtaining a return of less than 0.5 (i.e. 1 approval per 2 billion spent) is 79% for single drug portfolios, but only 7.6% for $N = 20$, 1.8% for $N = 40$, and 0.05% for $N = 100$.

Figure 11 provides a more detailed look at the risk-return characteristics of drug portfolios, as a function of portfolio size. The top left panel plots average portfolio returns (R_N) obtained from our simulations, for each N from 1 to 100. Mirroring the results from Figure 10, the plot shows a substantial increase in average returns as a function of portfolio size, up to a scale of about 20 drugs, after which they rapidly converge towards 1.06, which is the ratio of average number of approvals to average total costs. The top right panel of the figure shows the 25th and 75th percentiles of each return distribution, alongside the means. This panel shows us another view at the rapid decrease in return dispersion as a function of portfolio size, which substantially tapers off around portfolios of about 40 drugs. Consistent with this tapering behavior, while the bottom left panel of the figure shows that return standard deviations continue to decline beyond $N = 40$, the bottom right panel shows that they do so at a very slow pace, with each additional drug decreasing portfolio standard deviation by about 0.5%.

To gain a better understanding of the behavior of portfolio returns as a function of portfolio size, we perform a second order Taylor approximation of the return function around the expectation of number of approvals and expectation of development costs and obtain:

$$E(R_N) \equiv E(A_N/C_N) \approx \frac{E(A_N)}{E(C_N)}\left[1 - \frac{Cov(A_N, C_N)}{E(A_N)E(C_N)} + \frac{Var(C_N)}{E(C_N)^2}\right] \tag{5}$$

$$= \frac{E(A_N/N)}{E(C_N/N)}\left[1 - \frac{Cov(A_N/N, C_N/N)}{E(A_N/N)E(C_N/N)} + \frac{Var(C_N/N)}{E(C_N/N)^2}\right] \tag{6}$$

where A_N is the number of approvals in a portfolio of N drugs (i.e. the numerator in Equation 4), and C_N is the total cost of development in a portfolio of N drugs (i.e. the denominator in Equation 4).

Equation 6 shows that the expected return on a portfolio of drugs depends not only on the average rate of approval ($E(A_N/N)$) and average development costs per drug ($E(C_N/N)$), but also on the covariance between approval frequency and average development costs ($Cov(A_N/N, C_N/N)$) and the variance of average development costs ($Var(C_N/N)$). It can be shown that, if drug-development outcomes are pairwise independent, both the covariance term and the variance term converge to 0 as $N \to \infty$, and the expected return on the portfolio converges to the ratio of probability of success (PoS) to average development cost (i.e. $E(A_N/N)/E(C_N/N)$). However, if, as we document in Section 5, drug-development outcomes are correlated, these terms converge to a function of the average covariance between phase transitions of different drugs (and average development cost for each phase). The limiting behavior documented in the top left panel of Figure 11 suggests that, for moderately-sized portfolios of about 20 drugs, these pairwise correlations between drug-development outcomes have almost no effect on the average number of approvals per billion spent.

To examine how portfolio return standard deviation depends on portfolio size, we use a first order Taylor approximation of returns around the expectation of number of approvals and expectation of development costs and obtain:

$$\sigma(R_N) \equiv \sigma\left(\frac{A_N}{C_N}\right) \approx \frac{E(A_N)}{E(C_N)}\left[\frac{\text{Var}(A_N)}{E(A_N)^2} - 2\frac{\text{Cov}(A_N, C_N)}{E(A_N)E(C_N)} + \frac{\text{Var}(C_N)}{E(C_N)^2}\right]^{1/2} \tag{7}$$

$$= \frac{E(A_N/N)}{E(C_N/N)}\left[\frac{\text{Var}(A_N/N)}{E(A_N/N)^2} - 2\frac{\text{Cov}(A_N/N, C_N/N)}{E(A_N/N)E(C_N/N)} + \frac{\text{Var}(C_N/N)}{E(C_N/N)^2}\right]^{1/2} \tag{8}$$

Once again, it can be shown that, if drug-development outcomes are pairwise independent, the three terms in brackets in Equation 8 converge to 0 as $N \to \infty$, leading the standard deviation of portfolio returns to converge to 0. The bottom left panel in Figure 11 shows that, even for portfolios of 100 drugs, the return standard deviation is well above 0, at 0.18, and decreasing by 0.5% per additional drug. This suggests that cross-drug outcome correlations do play a significant role as a limiting factor for portfolio diversification.

6.2 The effects of increased correlations in drug-development outcomes

The fact that portfolios of 20 drugs have return standard deviations that are 70% lower than the average standard deviation of individual drugs means that the benefits of diversification are still substantial in drug development. However, it is important to recognize that this finding may overestimate the extent to which one can diversify away idiosyncratic risk in drug development in recent years. First, our simulations assume that all drugs in our sample are available to invest in at the same time, which underestimates potential cross-sectional correlation in drug development outcomes driven by common macroeconomic shocks. Second, our results in Section 5 show evidence of increased concentration of drug-development around similar targets, which leads to more correlation between drug-development outcomes in recent years.

In this section, we examine the extent to which the benefits of constructing portfolios of drug-development project has changed over time. To this end, we run our simulations separately on different cohorts of drugs which started development at similar times. To determine when each drug development programs started, we merge the Pharmaprojects database on drug-development programs (Section 3.3) with the trial-level data from Trialtrove (Section 3.1).[24] This allows us to identify trials associated with each drug-development program. When a phase 1 trial was not found for a program, we impute its starting date from the starting date of the earliest available clinical trial, by subtracting the average duration of the missing phases for that therapeutic area (see Table 2).

Each year from 1998 to 2013 we re-run the simulations described above using only the drugs that started development that year. We begin in 1998 because it is the first year in our sample with at least 100 targets. We end in 2013 to limit the potential for a downward bias in approval rates caused by the fact that we can only include completed drug-development programs in our tests. As Table 2 shows, this process takes an average of 6 years, and significantly higher for oncology drugs (about 10 years). Since our sample ends in 2023, looking at programs starting after 2013 will bias the sample towards programs that have completed significantly earlier than average (particularly

[24]The merge is done at the drug-disease-sponsor-country level. Matching on drugs is done using the "drugId" variable common to both datasets. Merging on disease is done through a combination of matching on disease names, and several standard disease classification codes: "meshId", "icd9Id", "icd10Id", and "snomedId". Merging by sponsor is done by matching sponsor "parentCompanyId" in Trialtrove, with "companyId" in Pharmaprojects (or with Pharmaprojects' "parentCompanyId" variable when "companyId" does not match).

for oncology drugs, which constitute over 30% of our sample), which is more likely to be the case for failed projects. Finally, to account for the well documented increase in clinical trial costs over time, we use the estimates from DiMasi, Hansen, and Grabowski (2003) to randomly draw trial costs for programs starting development prior to 2000. The authors document average out-of-pocket costs of \$15.2, \$23.5, and \$86.3 billion (in 2000 USD) and standard deviations of \$12.8, \$22.1, \$60.6 for phases 1, 2, and 3, respectively. For trials starting between 2000 and 2013, we linearly interpolate between these estimates and those from DiMasi, Grabowski, and Hansen (2016) discussed above (after converting both to 2023 USD).

Table 9 describes portfolio average returns obtained from running our simulations on a year-by-year basis as described above. Panel A presents average returns for portfolios of 1, 20, and 100 drugs ($E(R_1)$, $E(R_{20})$, and $E(R_{100})$ respectively) as well as the percentage improvement in average returns going from portfolios of 1 drug to portfolios of 20 drugs and 100 drugs. The first column in Panel A shows that average returns on portfolios of 1 drug were about 5 times larger in 1998 than in 2013. However, the second column shows that portfolios of 20 drugs experienced a much lower decline. The third column shows that the benefits to scale have significantly increased over time. For drugs starting development in 2013, portfolios of 20 drugs have average returns that are 115% larger than the average returns on single-drug portfolios. This improvement in average returns was only 14% in 1998. Finally, the last two columns in Panel A show that increasing portfolio size beyond 20 drugs has never provided any additional benefits in terms of increasing average portfolio returns.

Panel B in Table 9 provides a decomposition of average portfolio returns for $N = 20$ by examining how the terms in Equation 5 have changed over time. The key takeaway from this panel is that the decrease in portfolio average returns we document in the second column in Panel A has mainly been driven by a decrease in approval probability and increase in clinical trial costs. To see this, note that the first column in Panel B shows that the average number of approvals for a portfolio of 20 drugs has dramatically decreased over time. At the same time, the second column shows that average costs for such a portfolio have remained quite stable. While clinical trial costs have nearly doubled in this period (Scannell et al. (2012),DiMasi, Grabowski, and Hansen (2016)), the likelihood of reaching phase 3 (by far the most costly phase) has significantly decreased, resulting in stable average costs. The third and fourth columns show that both the variance of total costs

($Var(C_{20})$) and their covariance with the number of approvals ($Cov(A_{20}, C_{20})$) have increased over time. However, they enter Equation 5 with opposite sign. This causes the term in brackets in that equation to be nearly 1 every year (see "Multiplier" column). Hence, the decrease in portfolio returns is dominated by the decrease in average approvals to average costs ($E(A_{20})/E(C_{20})$).

Table 10 describes the evolution of return standard deviations for portfolios of 1, 20, and 100 drugs. The first, second, and fourth columns show significant decreases in standard deviations over time at all scales (we will explore this further in Panel B). The third column in Panel A shows substantial benefits to diversification every year. While the extent to which portfolios of 20 drugs are able to diversify away idiosyncratic risk has decreased over time, it remains quite significant (e.g. a 64% reduction for portfolios constructed in 2013). The last column in Panel A shows that portfolios of 100 drugs see only a minimal decrease in diversification benefits over time.

In Panel B of Table 10 we aim to gain a better understanding of the decrease in portfolio standard deviations over time, focusing on portfolios of 20 drugs. Since we are particularly interested in how the variance in total approvals contributes to the change in return standard deviations, we re-write Equation 7 as:

$$\sigma(R_N) \approx \left[\frac{\text{Var}(A_N)}{E(C_N)^2} - 2\frac{\text{Cov}(A_N, C_N)E(A_N)}{E(C_N)^3} + \frac{\text{Var}(C_N)E(A_N)^2}{E(C_N)^4} \right]^{1/2} \tag{9}$$

and we present how the three terms on the right hand side of this equation have evolved over time in the first three columns of Panel B. A comparison of these three columns reveals that the decrease in the $Var(A_20)/E(C_20)^2$ term has had the highest contribution to the decrease in portfolio return standard deviations over time. The last two columns of Panel B further break down this term and show that its decline is mainly driven by the fact that the variance in portfolio approvals ($Var(A_{20})$) has significantly *decreased* over time. It can be shown that, as N increases, this variance converges to the average pairwise covariance in drug approvals. The fact that this has decreased over time suggests that, while the increase in concentration around similar targets may have created larger pockets of highly correlated outcomes, this effect is dominated by the significant increase in the total number of targets, resulting in lower *average* pairwise correlations between drugs.

We conclude our analysis with an investigation of the extent to which poor diversification across biological targets can affect the risk-return characteristics of drug portfolios. To this end, we run

simulations where portfolios are built by randomly drawing from the set of available *targets* each year (as opposed to randomly drawing from the set of unique drugs). Once a target is selected, we include in the portfolio all the drugs with that target starting development that year. We repeat the process until we hit the desired total number of drugs in the portfolio.

The first three columns in Table 11 present the average returns, standard deviations, and ratio of average return to standard deviation obtained from this "target clusters" portfolio construction method. To ensure a clean comparison, the middle three columns present results obtained from our regular "random drugs" construction method, this time restricting ourselves to the drugs for which we have target data.[25] The last three columns show how portfolio characteristics change when moving from a "random drugs" approach to a "target clusters approach". These results show that improper diversification across targets results in higher portfolio standard deviations in every year of our sample, and even lower average returns in nine out of the 16 years. This amounts to significantly lower ratios of average returns to standard deviations, particularly in the second half of the sample. It is important to note that this "target clusters" approach is a much closer reflection of how pharmaceutical companies actually develop their portfolios of drugs given the specialized nature of expertise required to develop drugs for each different target. Our findings suggest that this portfolio construction approach does not take full advantage of the benefits of scale in drug development.

7 Conclusion

Our study documents an escalation in underinvestment risk in drug development, primarily driven by financial constraints and evolving sponsor dynamics. Over the past two decades, large pharmaceutical companies have notably reduced their participation in early-phase clinical trials, and this gap has been filled by smaller pharmaceutical firms and academic institutions, entities that often face significant financial limitations. Coupled with prolonged clinical trial durations and an increased focus on drugs targeting unproven biological mechanisms, the industry is experiencing heightened exposure to external financing fluctuations. We provide direct evidence of this phenomenon by showing that the proportion of clinical trials terminated due to lack of funding has

[25] A comparison of these columns with the $E(R_{20})$ column in Table 9 and the $\sigma(R_{20})$ column in Table 10 reveals almost identical results with the simulations using the full sample of drugs.

increased steadily over time, and significantly more after periods of low equity market sentiment, low IPO volume, or high equity market volatility. These trends raise concerns about the sustainability of current funding models and the potential abandonment of projects that hold substantial societal value due to financial constraints rather than clinical viability.

To address these challenges, we explored securitization as a potential financing solution that could mitigate financial risks associated with drug development. By aggregating multiple drug development projects into diversified portfolios, securitization can diminish idiosyncratic risks and offer a more attractive investment proposition. Our simulation analysis demonstrates that portfolios consisting of 20 drugs significantly reduce investment risk and increase average returns compared to single-drug investments. These findings suggest that portfolio-based financing approaches can improve capital efficiency and attract a broader range of investors to support pharmaceutical innovation.

However, our study also highlights the critical role of correlations between drug development projects in determining the effectiveness of diversification strategies. We document an increased concentration of drugs targeting similar biological mechanisms and show that it leads to correlated approval outcomes, potentially diminishing the benefits of diversification. Despite this trend, our simulations reveal that the expansion in the total number of unique biological targets over time has provided opportunities to construct portfolios with lower average correlations between projects, counteracting the effect of increased target concentration. Our findings show that, by carefully selecting drugs with diverse targets, investors can maximize risk reduction and improve the risk-reward tradeoffs associated with drug development projects.

References

Aghamolla C, Thakor RT, 2022. Do mandatory disclosure requirements for private firms increase the propensity of going public? Journal of Accounting Research 60(3):755–804.

Angus DC, Alexander BM, Berry S, Buxton M, Lewis R, et al., 2019. Adaptive platform trials: definition, design, conduct and reporting considerations. Nature Reviews Drug Discovery 18(10):797–807.

Baker, M., and J. Wurgler, 2006, Investor sentiment and the cross-section of stock returns, Journal of Finance 61, 1645-1680.

Besanko D, Thakor AV, 1987. Competitive equilibrium in the credit market under asymmetric information. Journal of Economic Theory 42(1):167–82.

Brown JR, Fazzari SM, Petersen BC, 2009. Financing innovation and growth: cash flow, external equity, and the 1990s R&D boom. Journal of Finance 64(1):151–85.

Chaudhuri SE, Cheng K, Lo AW, Pepke S, Rinaudo S, et al., 2019. A portfolio approach to accelerate therapeutic innovation in ovarian cancer. Journal of Investment Management 17(2):5–16.

Chaudhuri SE, Lo AW, 2020. Financially adaptive clinical trials via option pricing analysis. Journal of Econometrics, p.105026

Chen H, Lu D, Xiao Z, Li S, Zhang W, Luan X, Zhang W, Zheng G, 2024. Comprehensive applications of the artificial intelligence technology in new drug research and development. Health Information Science and Systems 12(1):41.

Das S, Rousseau R, Adamson PC, Lo AW, 2018. New business models to accelerate innovation in pediatric oncology therapeutics: a review. JAMA Oncology 4(9):1274–80.

DiMasi JA, Grabowski HG, Hansen RW, 2016. Innovation in the pharmaceutical industry: new estimates of R&D costs. Journal of Health Economics 47:20–33.

DiMasi JA, Hansen RW, Grabowski HG, 2003. The price of innovation: new estimates of drug development costs. Journal of Health Economics 22(2):151–85.

Fagnan DE, Gromatzky AA, Stein RM, Fernandez J-M, Lo AW, 2014. Financing drug discovery for orphan diseases. Drug Discovery Today 19(5):533–38.

Fagnan DE, Yang NN, McKew JC, Lo AW, 2015. Financing translation: analysis of the NCATS rare-diseases portfolio. Science Translational Medicine 7 (276):276ps3.

Fernandez JM, Stein RM, Lo AW, 2012. Commercializing biomedical research through securitization techniques. Nature Biotechnology 30:964–75.

Giambona E, Golec J, Lopez-de-Silanes F, 2021. Do firms purposefully change capital structure? Evidence from an investment-opportunity shock to drug firms. Journal of Financial and Quantitative Analysis 56(3):915–44.

Hall BH, Lerner J, 2010. The financing of R&D and innovation. In Handbook of the Economics of Innovation, Vol. 1, ed. BH Hall, N Rosenberg, pp. 609–39. Amsterdam: North-Holland.

Jorring A, Lo AW, Philipson TJ, Singh M, Thakor RT, 2021. Sharing R&D risk in healthcare via FDA hedges. Review of Corporate Finance Studies. In press. https://doi.org/10.1093/rcfs/cfab024.

Kerr WR, Nanda R, 2015. Financing innovation. Annual Review of Financial Economics 7:445–62.

Krieger J, Li D, Papanikolaou D, 2022. Missing novelty in drug development. Review of Financial Studies 35(2):636–79.

Kyle MK, 2018. Are important innovations rewarded? Evidence from pharmaceutical markets. Review of Industrial Organization 53(1):211–34.

Lo AW, 2017. Discussion: new directions for the FDA in the 21st century. Biostatistics 18(3):404–7.

Lo AW, Siah KW, 2021. Financing correlated drug development projects. Journal of Structured Finance 27(1):17–34. https://doi.org/10.3905/jsf.2020.1.114.

Lo, A.W. and Thakor, R.T., 2022. Financing biomedical innovation. Annual Review of Financial Economics, 14(1), pp.231-270.

Lorenzoni L, Dougherty S, 2022. Understanding Differences in Health Care Spending: A Comparative Study of Prices and Volumes Across OECD Countries. Health Services Insights 15:11786329221109755.

Mulcahy AW, Schwam D, Lovejoy SL, 2024. International Prescription Drug Price Comparisons: Estimates Using 2022 Data. Rand Health Quarterly 11(3):5.

Murphy, KM, Topel RH, 2006. The Value of Health and Longevity. Journal of Political Economy, 114, 871–904.

Myers SC, Majluf NS, 1984. Corporate financing and investment decisions when firms have information that investors do not have. Journal of Financial Economics 13(2):187–221.

Newey WK, West KD, 1987. Hypothesis testing with efficient method of moments estimation. International Economic Review 777–787.

Rampini AA, Viswanathan S, 2013. Collateral and capital structure. Journal of Financial Economics 109(2):466–92.

Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, et al. 2023. Artificial Intelligence and Machine Learning Technology Driven Modern Drug Discovery and Development. International Journal of Molecural Sciences 24(3):2026.

Scannell JW, Blanckley A, Boldon H, Warrington B, 2012. Diagnosing the decline in pharmaceutical R&D efficiency. Nature Reviews Drug Discovery 11(3):191–200.

Siah KW, Xu Q, Tanner K, Futer O, Frishkopf JJ, Lo AW, 2021. Accelerating glioblastoma therapeutics via venture philanthropy. Drug Discovery Today 26(7):1744–49.

Thakor, R.T., Anaya, N., Zhang, Y., Vilanilam, C., Siah, K.W., Wong, C.H. and Lo, A.W., 2017. Just how good an investment is the biopharmaceutical sector? Nature Biotechnology 35(12):1149–57.

Thakor RT, Lo AW, 2017. Optimal financing for R&D-intensive firms. NBER Working Paper 23831.

Wong CH, Siah KW, Lo AW, 2019. Estimation of clinical trial success rates and related parameters. Biostatistics 20(2):273–86. https://doi.org/10.1093/biostatistics/kxx069.

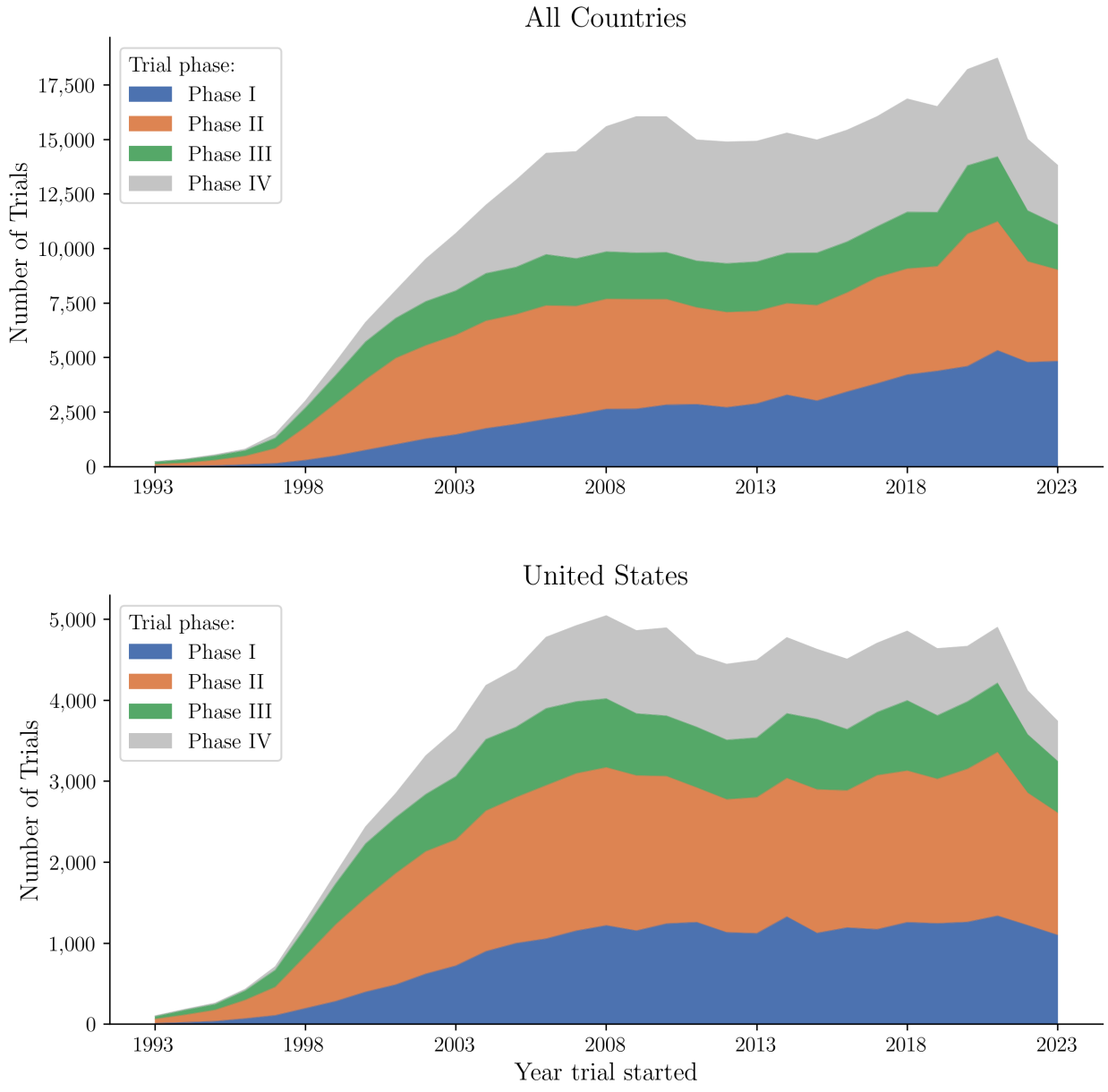

Figure 1
Number of trials by phase

This figure shows the number of trials initiated every year, by phase. The top panel includes trials from all countries in our dataset, while the bottom panel includes only trials in the United States.

Figure 2
Trial terminations

The upper left panel in this figure plots the percentage of trial terminations citing lack of funding and the upper right panel plots the percentage of terminations citing a business decision. The bottom panels break down these trends for trials with and without a top 20 pharma sponsor.

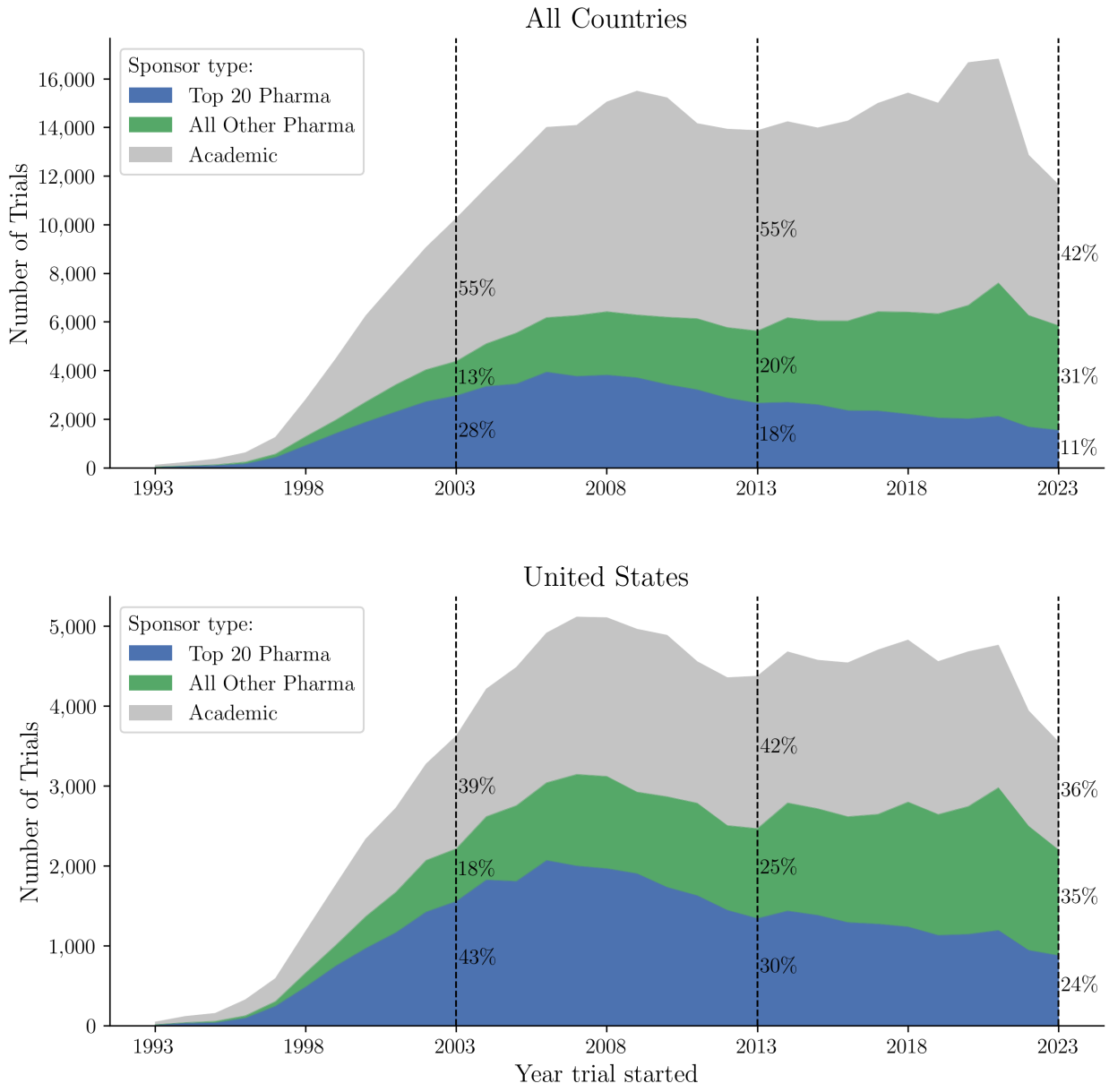

Figure 3
Number of trials by sponsor type

This figure shows the number of trials initiated every year, by sponsor type. The top panel includes all trials in our dataset, while the bottom panel includes only trials in the United States. The vertical dashed lines indicate the percentage of trials with sponsors from each category in those respective years.

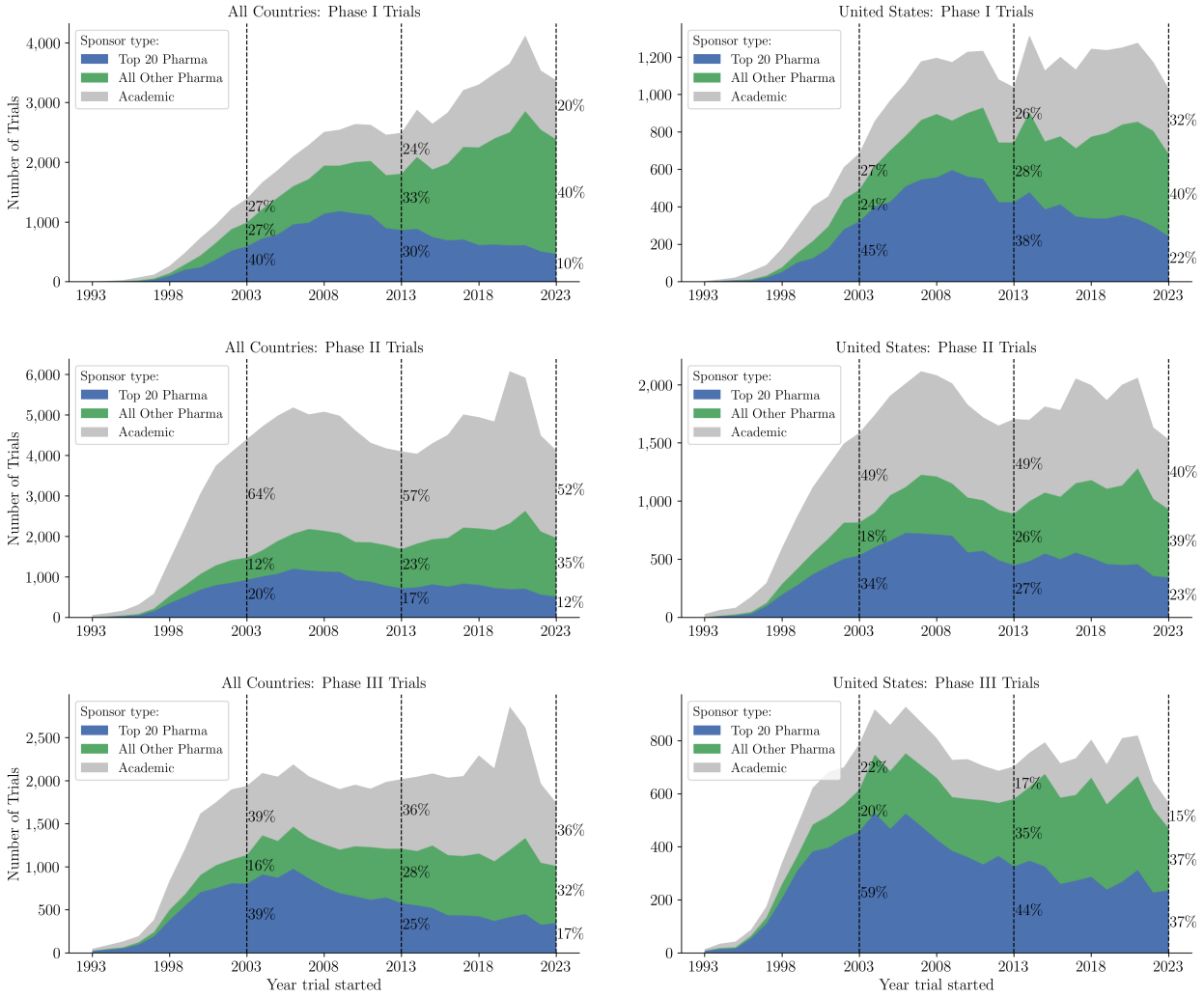

Figure 4
Number of trials by phase and sponsor type

This figure shows the number of trials initiated every year, by trial phase and sponsor type. The left panels include all trials in our dataset, while the right panels include only trials in the United States. The vertical dashed lines indicate the percentage of trials with sponsors from each category in those respective years.

Figure 5
Average trial duration from start of phase I to end of phase III

This figure shows the sum of average trial durations for Phases I, II, and III, for each therapeutic area in the Trialtrove dataset. Each panel corresponds to a different therapeutic area and presents averages for trials taking place in the United States as well as averages over all trials in the dataset.

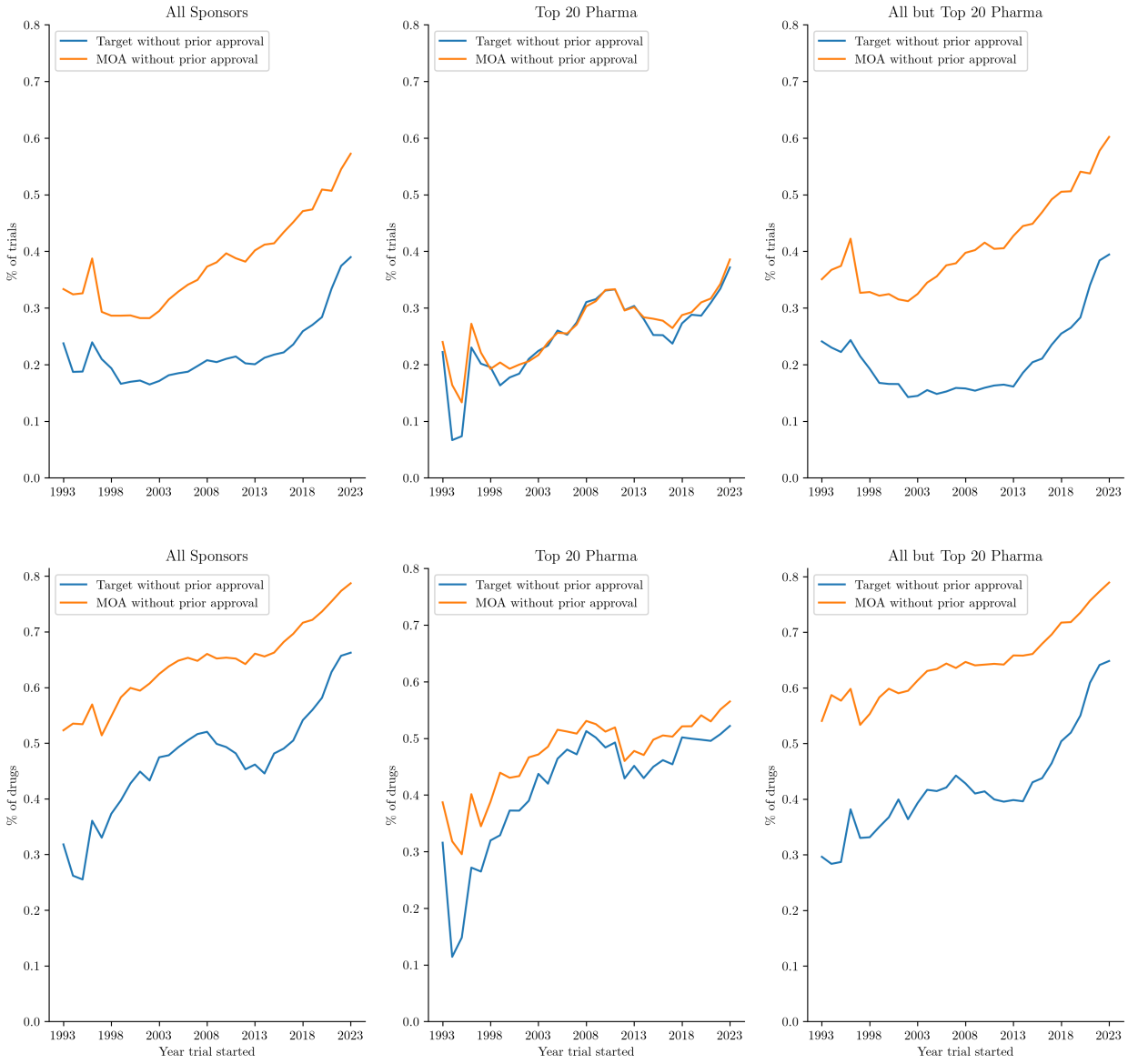

Figure 6
Trends in targets and MOAs with no prior approval

The top three panels in this figure show the percentage of trials investigating targets, or mechanisms of action (MOAs) with no prior approval. The bottom three panels show these same three categories as a percentage of the number of unique drugs investigated in trials starting each year. The leftmost panels in each row use all the trials in our dataset, the middle two use only trials sponsored by a top 20 pharma company, and the rightmost two use all the trials not sponsored by a top 20 pharma company.

N drugs = 47,546; N targets = 3,464; Drugs per target = 13.7

Gini index = 0.78

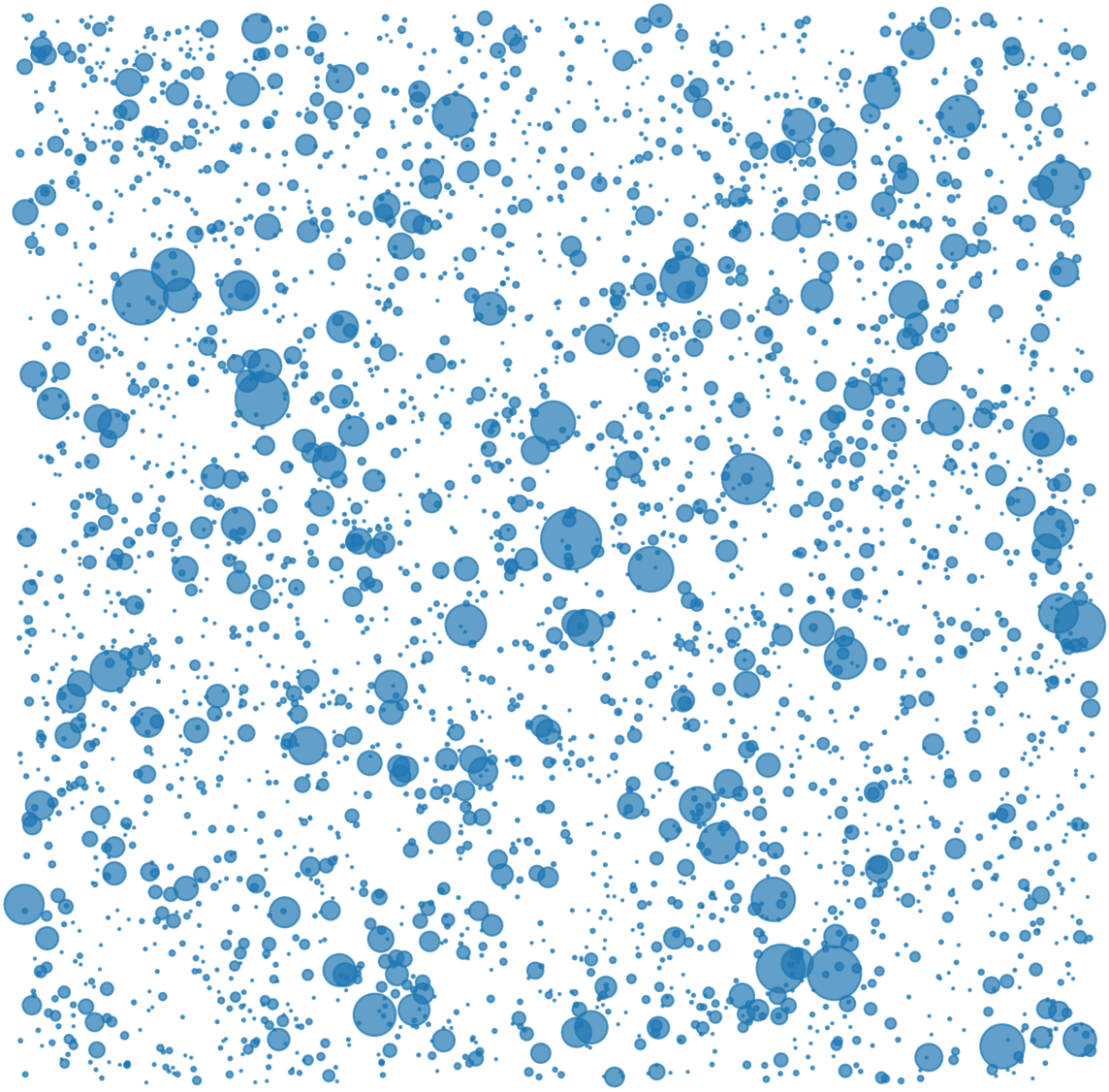

Figure 7
Drug-target concentration

This figure illustrates the concentration of drugs around the same target using the full Pharmaprojects database. Each node in the figure is a different biological target found in the database, and the size of each node is proportional to the number of drugs with that target in the sample.

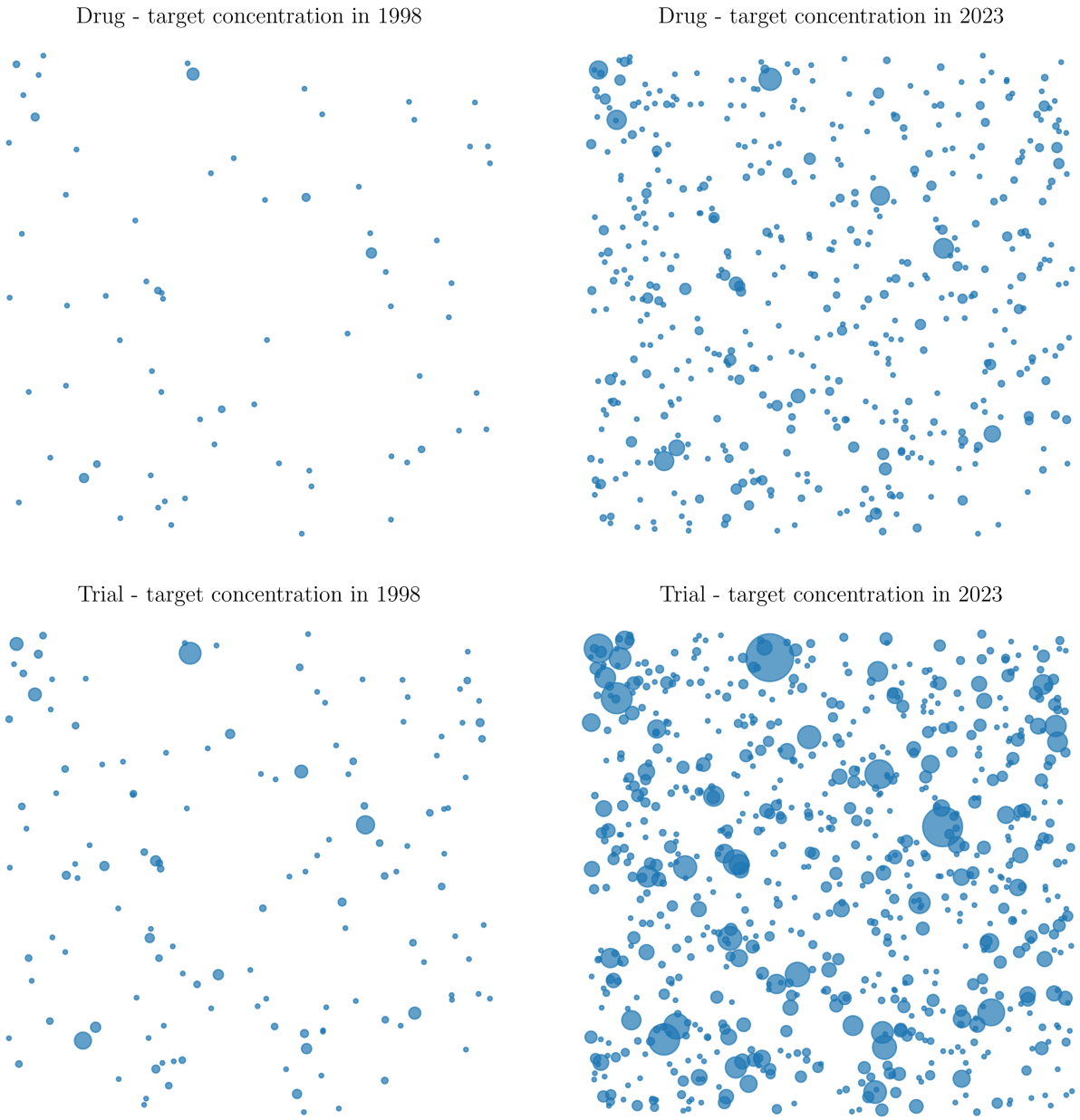

Figure 8
Evolution of concentration in target selection

This figure illustrates the concentration of drugs around the same target for drug development programs initiated in 1998 (leftmost panels) and 2023 (rightmost panels). Each node in the figure is a different biological target found in the database. In the top two panels, the size of each node is proportional to the number of drugs with that target and in the bottom two panels, the size of each node is proportional to the number of trials investigating drugs with that target.

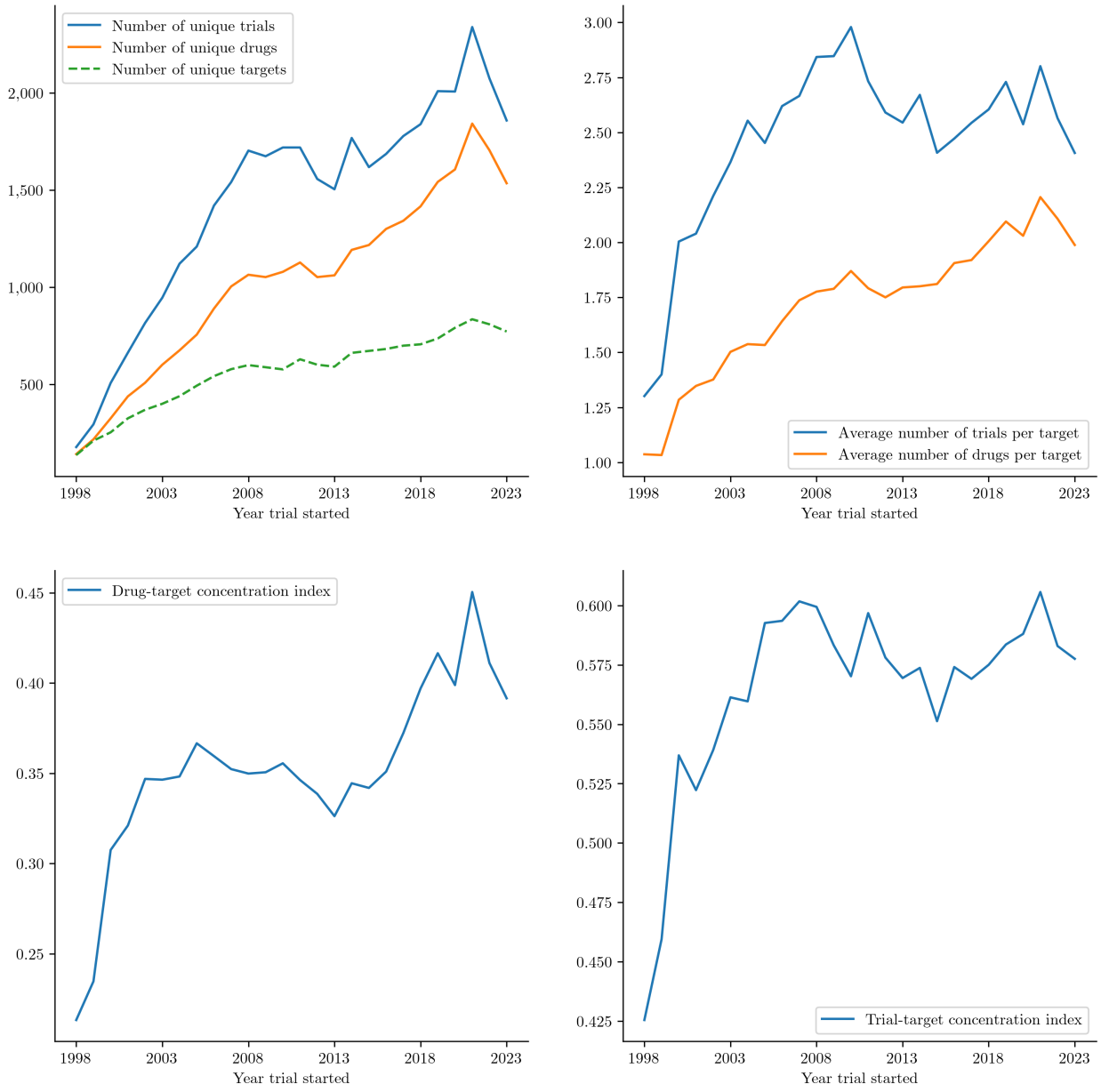

Figure 9
Trends in target concentration

The top left panel shows the evolution of number of unique trials, drugs, and targets over time. The top right panel shows the average number of trials per target and average number of drugs per target. The bottom panels plot the Gini index of inequality in the size of drug-target clusters (left) and trial-target clusters (right). In all panels, for each year, we use only the drug-development programs that started Phase I that year.

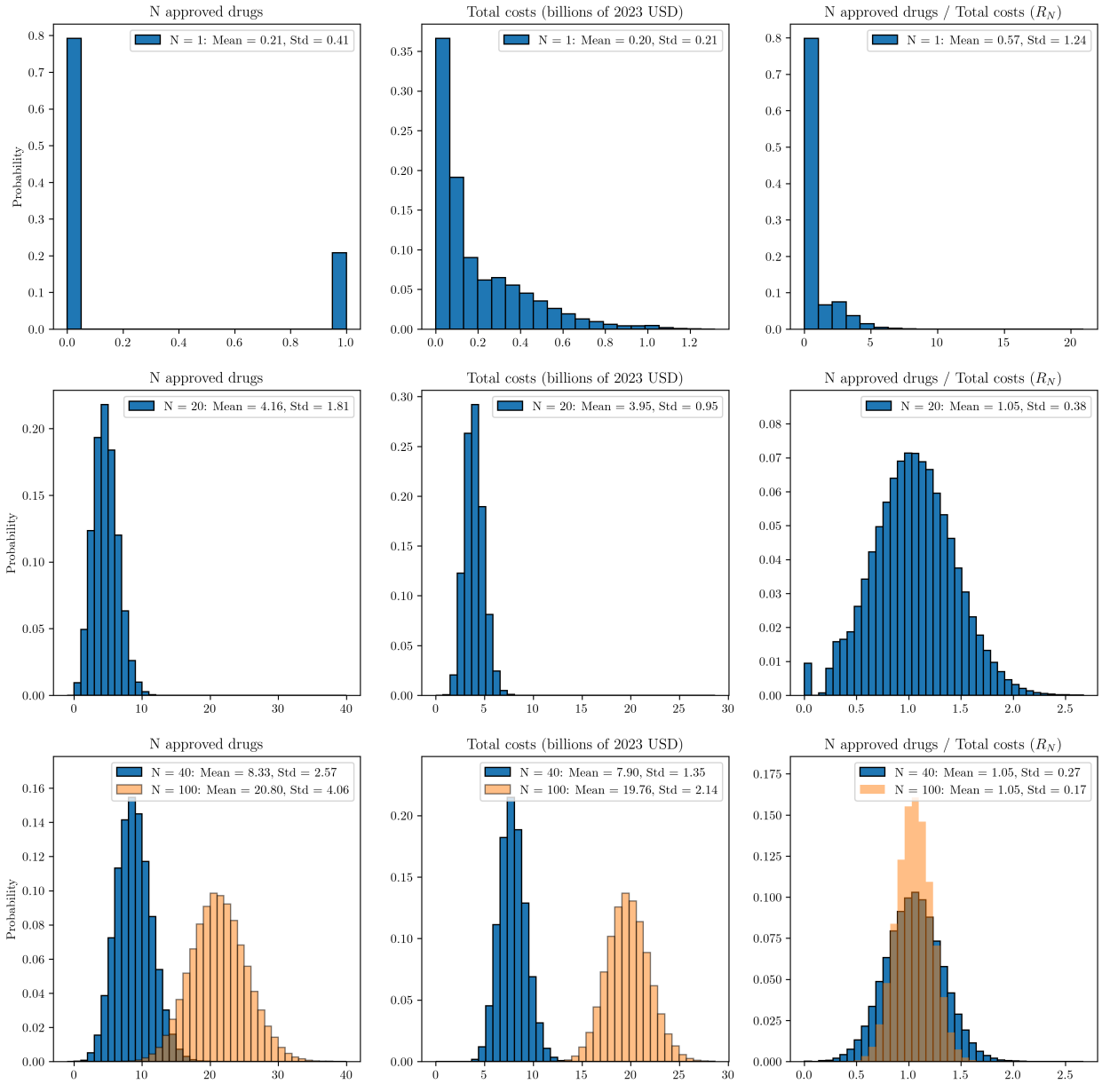

Figure 10
Simulated distribution of drug portfolio approvals, costs, and returns

This figure shows histograms of the simulated number of approved drugs per portfolio (left), total portfolio costs (middle) and the ratio of the two (right) for portfolios of 1 (top), 20, (center), 40 and 100 drugs (bottom). See Section 6.1 for a more detailed discussion on how these simulations are run.

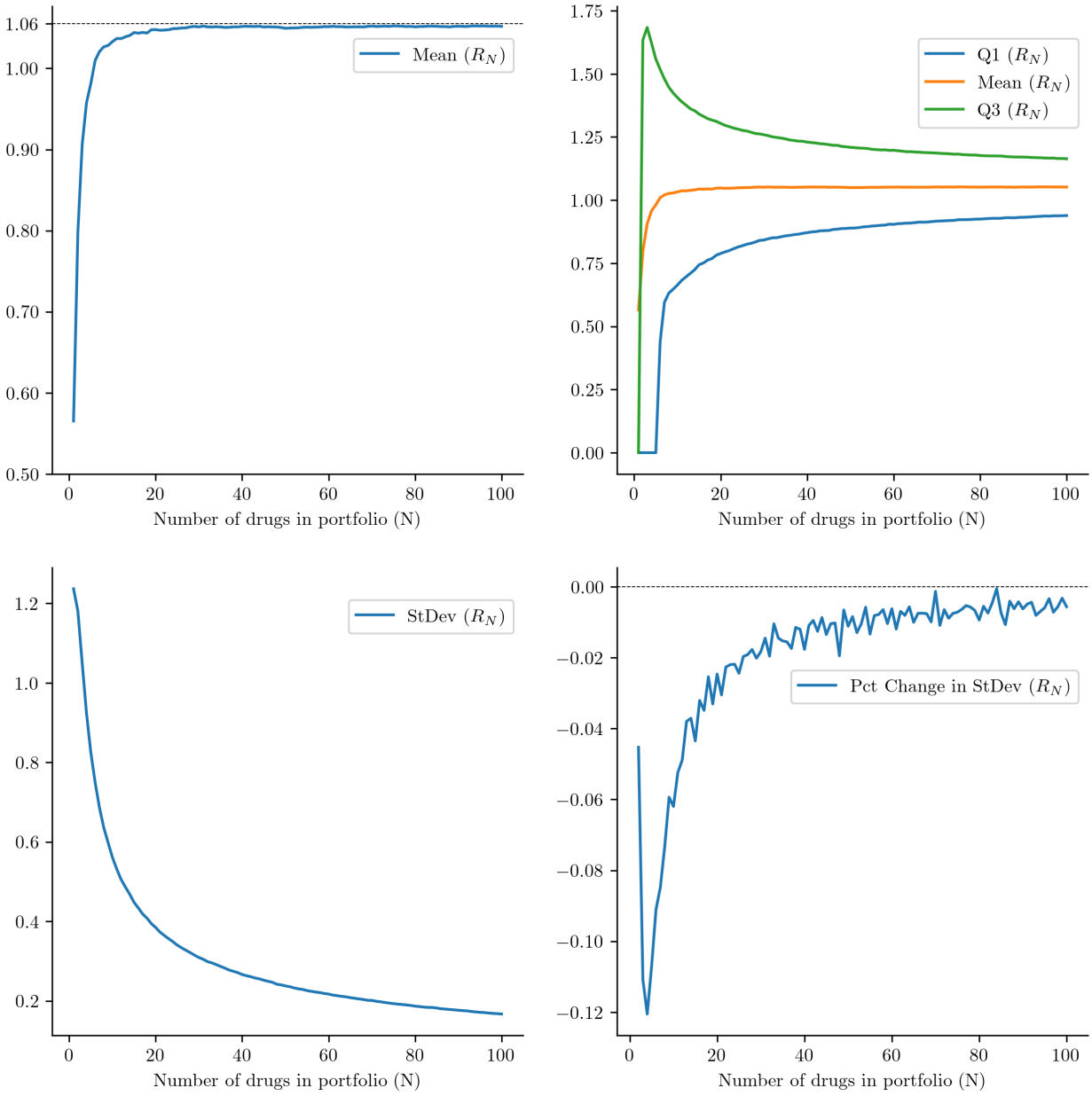

Figure 11
Portfolio average returns and standard deviations

This figure shows risk-return characteristics of drug portfolios, as a function of portfolio size. Using the simulated distributions of portfolio returns, we plot return averages in the top left panel, interquartile ranges in the top right panel, standard deviations in the bottom right panel, and the percentage change in standard deviations in the bottom right panel. See Section 6.1 for a more detailed discussion on how these simulations are run.

Table 1
Clinical Trial Summary Statistics

This table presents summary statistics for the trial-level data obtained from Trialtrove. Statistics are compiled for the full dataset (first two columns), the top 40 countries where the most trials were conducted (the middle two columns), and for the United States alone (last two columns). Panel A shows covarage statistics for some key trial characteristics. The remaining panels summarize the breakdown of trials by trial status (Panel B), trial phase (Panel C), sponsor types (Panel D) and main therapeutic area of the disease targeted by the trial (Panel E). In panels A and B, the percentages are calculated based on the topline "Trials" count (the first row in the table). In panels C, D, and E, the percentages are calculated based on the number of trials with available information on the respective variable. For panels D and E, these percentages may add up to more than 100% as a trial can have multiple sponsors or target multiple therapeutic areas.

	All 200 countries		Top 40 Countries		United States	
	N	%	N	%	N	%
Panel A: Data coverage						
Trials	407,935	1.00	374,191	1.00	123,656	1.00
Trials with phase data	380,446	0.93	349,649	0.93	117,231	0.95
Trials with sponsor data	371,642	0.91	342,891	0.92	111,565	0.90
Trials with therapeutic area data	358,511	0.88	331,637	0.89	104,419	0.84
Panel B: Trial status						
Ongoing	58,530	0.14	55,737	0.15	16,718	0.14
Completed	303,032	0.74	278,952	0.75	86,173	0.70
Terminated	46,373	0.11	39,502	0.11	20,765	0.17
Panel C: Trial phase						
Phase I	81,788	0.21	71,769	0.21	28,187	0.24
Phase II	127,427	0.33	119,116	0.34	47,933	0.41
Phase III	61,947	0.16	57,402	0.16	21,554	0.18
Phase IV	109,284	0.29	101,362	0.29	19,557	0.17
Panel D: Sponsor types						
Academic	206,831	0.56	195,719	0.57	48,948	0.44
Top 20 Pharma	74,356	0.20	68,227	0.20	38,270	0.34
All Other Pharma	84,962	0.23	75,675	0.22	31,308	0.28
Government	32,316	0.09	30,801	0.09	18,146	0.16
Panel E: Therapeutic areas						
Oncology	90,991	0.25	87,404	0.26	35,707	0.34
CNS	66,406	0.19	61,561	0.19	21,568	0.21
Autoimmune/Inflammation	51,561	0.14	47,940	0.14	13,657	0.13
Infectious Disease	48,596	0.14	42,999	0.13	11,687	0.11
Cardiovascular	47,389	0.13	43,790	0.13	9,404	0.09
Metabolic/Endocrinology	47,363	0.13	42,498	0.13	11,822	0.11
Vaccines (Infectious Disease)	11,467	0.03	9,929	0.03	2,936	0.03
Genitourinary	11,261	0.03	10,290	0.03	2,099	0.02
Ophthalmology	9,375	0.03	8,725	0.03	2,491	0.02

Table 2
Average Trial Duration by Therapeutic Area

This table presents average trial durations (in years) by therapeutic area. Panel A uses the full Trialtrove database, while Panel B uses only trials taking place in the United States. Trial durations are calculated as the difference between the trial start and end dates, divided by 365.

	Phase I	Phase II	Phase III	I + II + III
Panel A: All countries				
Oncology	2.45	3.38	4.29	10.12
Cardiovascular	0.58	2.20	2.51	5.28
Ophthalmology	1.26	1.82	2.00	5.09
CNS	0.80	2.07	2.18	5.05
Metabolic/Endocrinology	0.63	2.06	2.13	4.82
Autoimmune/Inflammation	0.82	1.89	2.01	4.72
Genitourinary	0.68	1.80	2.03	4.50
Infectious Disease	0.84	1.56	1.73	4.14
Vaccines (Infectious Disease)	1.36	1.38	1.26	3.99
All therapeutic areas	1.19	2.49	2.37	6.05
Panel B: United States				
Oncology	2.95	3.56	4.11	10.63
Cardiovascular	1.06	2.30	2.74	6.10
CNS	1.17	2.27	2.36	5.80
Metabolic/Endocrinology	0.94	2.30	2.38	5.61
Autoimmune/Inflammation	1.25	2.17	2.14	5.56
Genitourinary	1.09	2.22	2.21	5.53
Ophthalmology	1.55	1.76	2.08	5.38
Infectious Disease	1.21	1.86	2.01	5.08
Vaccines (Infectious Disease)	1.72	1.69	1.51	4.93
All therapeutic areas	1.86	2.76	2.55	7.18

Table 3
Drug Summary Statistics

This table presents summary statistics for the drug-level data obtained from Pharmaprojects. Statistics are compiled for the full dataset (first two columns), and for the drugs with trial information (last two columns). Panel A shows covarage statistics for some key drug characteristics. The remaining panels summarize the breakdown of drugs by origin (Panel B), and main indication group (therapeutic area) of the disease targeted by the drug (Panel C). In panels A and B, the percentages are calculated based on the topline "Drugs" count (the first row in the table). In panels B and C, the percentages are calculated based on the number of drugs with available information on the respective variable. For panel C, these percentages may add up to more than 100% as a drug can target multiple therapeutic areas.

	All drugs		Drugs with trial data	
	N	%	N	%
Panel A: Data coverage				
Drugs	104,129	1.00	28,606	1.00
Drugs with origin data	102,744	0.99	28,344	0.99
Drugs with indication group data	101,228	0.97	28,392	0.99
Drugs with mechanism of action data	72,971	0.70	23,677	0.83
Drugs with target data	47,537	0.46	17,966	0.63
Panel B: Drug origin				
Biological	33,238	0.32	10,680	0.38
Chemical	66,133	0.64	16,864	0.59
Panel C: Drug indication groups				
Anticancer	30,653	0.30	9,302	0.33
Neurological	17,381	0.17	4,984	0.18
Anti-infective	16,687	0.16	4,518	0.16
Alimentary/Metabolic	11,547	0.11	4,071	0.14
Cardiovascular	8,921	0.09	2,606	0.09
Musculoskeletal	7,720	0.08	1,919	0.07
Respiratory	5,452	0.05	2,029	0.07
Immunological	5,362	0.05	1,916	0.07
Dermatological	4,465	0.04	1,719	0.06
Blood and Clotting	4,126	0.04	1,480	0.05
Sensory	3,719	0.04	1,345	0.05
Genitourinary (including sex hormones)	3,435	0.03	1,444	0.05
Miscellaneous	1,419	0.01	618	0.02
Hormonal (excluding sex hormones)	879	0.01	246	0.01
Antiparasitic	876	0.01	140	0.00

Table 4
Drug Development Programs

This table presents summary statistics for the Pharmaprojects data on drug-development programs. Statistics are compiled for the full dataset (first two columns), and for the United States alone (last two columns). Panel A shows a count of drug-development programs at several aggregation levels. Panels B and C break down the unique drug-disease-sponsor-country records by the current development status of each program and the highest phased they reached in the approval process.

	All countries		United States	
	N	%	N	%
Panel A: Number of drug development programs				
Unique drug-disease-sponsor-country records	584,726		114,450	
Unique drug-disease-sponsor records	226,356		114,450	
Unique drug-disease records	174,937		92,671	
Panel B: Development status				
Ceased	320,848	0.55	88,979	0.78
Widely Launched	166,780	0.29	6,799	0.06
Active	97,098	0.17	18,672	0.16
Panel C: Highest phase reached				
Preclinical	181,500	0.31	65,872	0.58
Phase I	45,619	0.08	14,565	0.13
Phase II	80,284	0.14	18,958	0.17
Phase III	88,077	0.15	7,035	0.06
Approved	189,246	0.32	8,020	0.07

Table 5
Phase Transition Probabilities

This table presents phase transition probabilities by therapeutic area for drug-development programs in the Pharmaprojects dataset that are not in active development. The probability of transitioning from each phase of development is calculated as 1 minus the number of drugs that fail in that phase (as given by the "highestPhaseReached" variable in Pharmaprojects) divided by the total number of drugs that reached that phase. See Section 3.3 for more details. The second to last column shows the probability of approval starting from the preclinical phase, and is computed as the product of all four phase transition probabilities. The last column shows the likelihood of approval starting from phase 1, and is calculated as the product of $\mathbb{P}_{1-2}$, $\mathbb{P}_{2-3}$, and $\mathbb{P}_{3-A}$. In Panel A, a drug-development program is considered to have reached a given phase of development if it has done so in any country from the European Union, United States or Japan. In Panel B, only phase transitions from the United States are considered.

	N	$\mathbb{P}_{Pre-1}$	$\mathbb{P}_{1-2}$	$\mathbb{P}_{2-3}$	$\mathbb{P}_{3-A}$	$\mathbb{P}_{Pre-A}$	$\mathbb{P}_{1-A}$
Panel A: All countries							
Alimentary/Metabolic	10,146	40.5	70.3	48.3	64.8	8.9	22.0
Anti-infective	15,279	30.9	71.9	60.5	74.1	10.0	32.2
Anticancer	31,241	44.1	63.8	29.4	44.8	3.7	8.4
Antiparasitic	697	24.8	66.5	59.1	76.5	7.5	30.1
Blood and Clotting	3,587	42.3	73.8	61.3	71.4	13.7	32.3
Cardiovascular	8,738	41.0	76.9	55.4	63.4	11.1	27.0
Dermatological	3,789	51.3	77.1	46.7	73.1	13.5	26.3
Genitourinary (including sex hormones)	2,930	54.1	78.9	57.4	67.6	16.6	30.6
Hormonal (excluding sex hormones)	754	41.2	68.8	61.2	80.2	13.9	33.8
Immunological	4,340	38.1	69.5	49.2	66.6	8.7	22.8
Miscellaneous	1,044	46.5	78.8	63.6	60.9	14.2	30.5
Musculoskeletal	6,204	28.2	72.5	62.1	67.3	8.5	30.3
Neurological	17,494	37.7	71.6	48.4	60.5	7.9	21.0
Respiratory	5,328	45.5	74.8	46.9	64.0	10.2	22.5
Sensory	3,104	39.6	84.2	53.9	65.6	11.8	29.8
Total	114,675	40.0	70.6	46.8	63.0	8.3	20.8
Panel B: United States							
Alimentary/Metabolic	6,529	40.5	71.7	47.8	57.7	8.0	19.8
Anti-infective	10,085	30.8	69.2	57.2	69.7	8.5	27.6
Anticancer	23,254	48.2	62.7	28.3	39.8	3.4	7.1
Antiparasitic	423	27.9	74.6	58.0	78.4	9.5	33.9
Blood and Clotting	2,220	43.6	74.9	58.1	62.1	11.8	27.0
Cardiovascular	5,096	40.3	79.1	53.9	55.7	9.6	23.7
Dermatological	2,424	49.3	76.9	45.2	67.5	11.6	23.5
Genitourinary (including sex hormones)	1,859	53.8	78.7	58.9	64.4	16.1	29.9
Hormonal (excluding sex hormones)	452	41.2	71.0	62.9	78.3	14.4	34.9
Immunological	3,017	38.8	73.1	48.3	59.4	8.2	21.0
Miscellaneous	811	46.2	78.4	60.9	58.7	12.9	28.0
Musculoskeletal	3,776	27.4	70.8	58.2	57.5	6.5	23.7
Neurological	11,506	37.7	73.8	47.6	55.8	7.4	19.6
Respiratory	2,926	44.5	75.7	44.1	58.5	8.7	19.5
Sensory	2,221	40.3	85.2	49.3	61.2	10.4	25.7
Total	76,599	41.3	70.2	44.1	56.5	7.2	17.5

Table 6
Trial Outcomes

This table presents trial outcomes statistics for the Trialtrove data. Statistics are compiled for the full dataset (first two columns), and for the United States alone (last two columns). Panel A shows the different outcomes for completed trials and the percentages are calculated based on the number of completed trials with outcome data. Panel B present outcomes for terminated trials, and the percentages are based on the number of terminated trials with outcome data. In both panels, the percentages can add up to more than 1 since trials may have multiple outcomes listed.

	All countries		United States	
	N	%	N	%
Panel A: Completed trials				
Completed, Positive outcome/primary endpoint(s) met	32,832	0.45	14,109	0.48
Completed, Outcome unknown	24,097	0.33	6,350	0.22
Completed, Outcome indeterminate	8,202	0.11	5,163	0.18
Completed, Negative outcome/primary endpoint(s) not met	6,786	0.09	3,547	0.12
Completed, Early positive outcome	441	0.01	241	0.01
Panel B: Terminated trials				
Terminated, Planned but never initiated	11,956	0.36	2,936	0.19
Terminated, Unknown	10,626	0.32	2,607	0.17
Terminated, Poor enrollment	6,038	0.18	3,623	0.24
Terminated, Other	4,029	0.12	2,164	0.14
Terminated, Business decision - Other	3,917	0.12	2,259	0.15
Terminated, Business decision - Pipeline reprioritization	3,220	0.10	1,204	0.08
Terminated, Lack of efficacy	2,818	0.08	1,875	0.12
Terminated, Safety/adverse effects	1,623	0.05	1,021	0.07
Terminated, Lack of funding	1,268	0.04	902	0.06
Terminated, Business decision - Drug strategy shift	1,237	0.04	542	0.04

Table 7
Trial terminations due to funding

This table presents regression results on the determinants of trial terminations due to funding. In Panel A, we present trial-level regressions where the dependent variable is a binary indicator for whether the trial was terminated due to funding. Standard errors are clustered by year of trial termination. In Panel B, we present macro-level regressions where the dependent variable is the percentage of terminations due to funding in a given year. Standard errors are adjusted for serial correlation up to two lags, using the Newey-West (1997) estimator. Both panels include trials ending between 2000 and 2023. t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Trial-level regressions				
Top 20 Pharma	-0.04***	-0.04***	-0.04***	-0.04***
	(-5.11)	(-5.21)	(-5.31)	(-5.22)
Trial duration		0.03***	0.03***	0.03***
		(3.47)	(3.40)	(3.39)
Baker-Wurgler sentiment			-0.02***	-0.02***
			(-3.53)	(-4.07)
VIX				0.01***
				(3.21)
Linear trend	0.02***	0.02***	0.02***	0.02***
	(3.45)	(3.87)	(4.59)	(4.24)
Intercept	0.19***	0.14***	0.14***	0.10***
	(14.51)	(10.89)	(9.26)	(4.71)
R^2	0.002	0.003	0.003	0.003
Observations	25,070	25,070	25,070	25,070
Panel B: Macro-level regressions				
Linear trend	0.74***	0.70***	0.82***	0.82***
	(5.87)	(6.91)	(7.73)	(9.75)
Baker-Wurgler sentiment		-0.37***		-0.29***
		(-2.84)		(-3.43)
Aggregate IPO volume			-0.31**	-0.21**
			(-2.01)	(-2.28)
VIX				0.35***
				(7.77)
Intercept	2.22***	2.38***	2.49***	1.22***
	(7.58)	(9.22)	(9.67)	(5.27)
R^2	0.545	0.678	0.634	0.829
Observations	24	24	24	24

Table 8
Predicting Drug Approvals

This table presents results from regressing drug-program approval indicators on the average approval rate of other drugs with the same target and various drug-level characteristics. The sample includes only targets that have been developed by at least 2 drugs. Therapeutic areas are as defined by Pharmaprojects (i.e. indication groups as illustrated in Table 3). See Section 5 for a definition of target family and MOA family. Standard errors are clustered at the therapeutic-area level. t-statistics are reported in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

	1	2	3	4
Approval rate of other drugs with same target (AAR)	0.94***	0.92***	0.91***	0.89***
	(38.41)	(30.11)	(29.61)	(29.61)
Intercept	0.01*	0.07***	0.07***	0.08***
	(1.81)	(11.97)	(9.98)	(9.55)
R^2	0.205	0.208	0.209	0.212
Observations	30,546	30,546	30,546	30,546
Therapeutic area indicators	No	Yes	Yes	Yes
Target family indicators	No	No	Yes	Yes
MOA family indicators	No	No	No	Yes

Table 9
Evolution of drug portfolio average returns

This table describes average returns obtained from constructing simulated drug portfolios on a year-by-year basis as described in Section 6.2. Panel A presents average returns for portfolios of 1, 20, and 100 drugs ($E(R_1)$, $E(R_{20})$, and $E(R_{100})$ respectively) as well as the percentage improvement in average returns going from portfolios of 1 drug to portfolios of 20 drugs and 100 drugs (third and fifth column). Panel B provides a decomposition of average returns on portfolios of 20 drugs by examining how the terms in Equation 5 have changed over time. The "Multiplier" column calculates the term that appears in brackets inEquation 5.

Panel A: Drug portfolio average returns

	$E(R_1)$	$E(R_{20})$	$\frac{E(R_{20})}{E(R_1)} - 1$	$E(R_{100})$	$\frac{E(R_{100})}{E(R_1)} - 1$
1998	2.24	2.55	0.14	2.54	0.13
1999	1.85	2.25	0.22	2.24	0.22
2000	1.59	1.82	0.14	1.81	0.13
2001	1.50	1.98	0.32	1.98	0.32
2002	1.02	1.49	0.45	1.49	0.46
2003	0.70	1.18	0.67	1.19	0.69
2004	0.72	1.23	0.71	1.24	0.73
2005	0.47	0.91	0.93	0.93	0.97
2006	0.47	0.97	1.04	0.99	1.08
2007	0.50	1.07	1.13	1.09	1.17
2008	0.42	0.97	1.29	0.99	1.35
2009	0.37	0.89	1.39	0.91	1.46
2010	0.50	1.06	1.13	1.08	1.17
2011	0.48	1.09	1.28	1.11	1.32
2012	0.44	0.94	1.12	0.96	1.15
2013	0.46	1.00	1.15	1.01	1.18

Panel B: Decomposition of average returns

	$E(A_{20})$	$E(C_{20})$	$Cov(A_{20}, C_{20})$	$Var(C_{20})$	$\frac{E(A_{20})}{E(C_{20})}$	$Multiplier$
1998	8.02	3.16	0.50	0.26	2.54	1.01
1999	6.62	2.96	0.48	0.26	2.24	1.01
2000	5.68	3.15	0.36	0.26	1.81	1.01
2001	6.06	3.06	0.52	0.26	1.98	1.00
2002	4.52	3.03	0.47	0.28	1.49	1.00
2003	3.41	2.86	0.44	0.27	1.19	0.99
2004	3.72	2.99	0.52	0.32	1.24	0.99
2005	2.59	2.78	0.44	0.32	0.93	0.98
2006	2.77	2.79	0.51	0.34	0.99	0.98
2007	3.00	2.74	0.61	0.37	1.10	0.98
2008	2.68	2.68	0.61	0.41	1.00	0.97
2009	2.45	2.68	0.59	0.46	0.91	0.97
2010	3.32	3.07	0.80	0.55	1.08	0.98
2011	3.35	3.00	0.87	0.61	1.12	0.98
2012	3.17	3.30	0.83	0.68	0.96	0.98
2013	3.41	3.36	0.95	0.77	1.01	0.99

Table 10
Evolution of drug portfolio standard deviations

This table describes the standard deviation of returns obtained from constructing simulated drug portfolios on a year-by-year basis as described in Section 6.2. Panel A presents standard deviations of returns for portfolios of 1, 20, and 100 drugs ($\sigma(R_1)$, $\sigma(R_{20})$, and $\sigma(R_{100})$ respectively) as well as the percentage improvement in standard deviations going from portfolios of 1 drug to portfolios of 20 drugs and 100 drugs (third and fifth column). Panel B provides a decomposition of return standard deviations for portfolios of 20 drugs by examining how the terms in Equation 9 have changed over time.

Panel A: Drug portfolio return standard deviations

	$\sigma(R_1)$	$\sigma(R_{20})$	$\frac{\sigma(R_{20})}{\sigma(R_1)} - 1$	$\sigma(R_{100})$	$\frac{\sigma(R_{100})}{\sigma(R_1)} - 1$
1998	3.21	0.64	-0.80	0.28	-0.91
1999	3.05	0.65	-0.79	0.29	-0.91
2000	2.91	0.61	-0.79	0.27	-0.91
2001	2.55	0.60	-0.77	0.26	-0.90
2002	2.06	0.55	-0.73	0.24	-0.88
2003	1.67	0.52	-0.69	0.23	-0.86
2004	1.61	0.51	-0.69	0.22	-0.86
2005	1.30	0.48	-0.63	0.21	-0.84
2006	1.27	0.47	-0.63	0.21	-0.83
2007	1.27	0.48	-0.62	0.21	-0.83
2008	1.16	0.47	-0.60	0.21	-0.82
2009	1.08	0.46	-0.57	0.20	-0.81
2010	1.21	0.44	-0.64	0.19	-0.84
2011	1.18	0.44	-0.63	0.19	-0.84
2012	1.13	0.41	-0.64	0.18	-0.84
2013	1.14	0.41	-0.64	0.18	-0.85

Panel B: Decomposition of return standard deviations

	$\frac{\text{Var}(A_{20})}{E(C_{20})^2}$	$\frac{\text{Cov}(A_{20},C_{20})E(A_{20})}{E(C_{20})^3}$	$\frac{\text{Var}(C_{20})E(A_{20})^2}{E(C_{20})^4}$	$Var(A_{20})$	$E(C_{20})$
1998	0.48	0.13	0.17	4.79	3.16
1999	0.51	0.12	0.15	4.43	2.96
2000	0.41	0.07	0.09	4.03	3.15
2001	0.45	0.11	0.11	4.24	3.06
2002	0.38	0.08	0.07	3.51	3.03
2003	0.35	0.06	0.05	2.84	2.86
2004	0.34	0.07	0.05	3.04	2.99
2005	0.29	0.05	0.04	2.26	2.78
2006	0.31	0.07	0.04	2.38	2.79
2007	0.34	0.09	0.06	2.55	2.74
2008	0.32	0.08	0.06	2.33	2.68
2009	0.30	0.08	0.05	2.15	2.68
2010	0.29	0.09	0.07	2.77	3.07
2011	0.31	0.11	0.08	2.78	3.00
2012	0.25	0.07	0.06	2.68	3.30
2013	0.25	0.08	0.07	2.85	3.36

Table 11
Effects of target clustering on drug portfolio returns

This table presents results obtained from building simulated portfolios by randomly drawing from the set of available *targets* each year. Once a target is selected, we include all the drugs with that target starting development that year, and we continue the process until we have selected 20 drugs. The first three columns in this table present the average returns, standard deviations, and ratios of average returns to standard deviations obtained from this simulation strategy. The middle three columns show the same metrics, obtained from our regular simulation methodology, restricting ourselves only to drugs that have target data. The last three columns show how these metrics change going from the regular "random drugs" strategy to the new "target clusters" strategy.

	Using target clusters			Using random drugs			Percent difference		
	$E(R_t)$	$\sigma(R_t)$	$\frac{E(R_t)}{\sigma(R_t)}$	$E(R_d)$	$\sigma(R_d)$	$\frac{E(R_d)}{\sigma(R_d)}$	$\frac{E(R_t)}{E(R_d)}-1$	$\frac{\sigma(R_t)}{\sigma(R_d)}-1$	$\frac{E(R_t)/\sigma(R_t)}{E(R_d)/\sigma(R_d)}-1$
1998	2.60	0.80	3.27	2.42	0.64	3.78	0.08	0.24	-0.14
1999	2.70	0.83	3.26	2.36	0.65	3.61	0.14	0.27	-0.10
2000	2.16	0.78	2.77	1.89	0.62	3.04	0.14	0.26	-0.09
2001	1.90	0.90	2.10	1.86	0.59	3.13	0.02	0.52	-0.33
2002	1.51	0.68	2.24	1.45	0.54	2.68	0.04	0.24	-0.16
2003	1.12	0.70	1.61	1.20	0.53	2.28	-0.07	0.32	-0.30
2004	1.10	0.64	1.71	1.20	0.50	2.41	-0.09	0.29	-0.29
2005	1.02	0.53	1.93	0.95	0.47	1.99	0.07	0.11	-0.03
2006	1.02	0.53	1.94	0.91	0.47	1.95	0.13	0.13	-0.01
2007	0.95	0.63	1.52	1.06	0.48	2.23	-0.10	0.32	-0.32
2008	0.85	0.61	1.39	0.95	0.47	2.00	-0.10	0.29	-0.31
2009	0.82	0.58	1.41	0.91	0.46	1.97	-0.10	0.25	-0.28
2010	0.98	0.53	1.86	1.10	0.44	2.52	-0.11	0.21	-0.26
2011	1.13	0.55	2.06	1.14	0.43	2.63	-0.01	0.26	-0.22
2012	0.87	0.45	1.93	0.98	0.40	2.42	-0.12	0.11	-0.20
2013	0.87	0.46	1.88	1.01	0.41	2.46	-0.14	0.13	-0.24

Appendix A

Figure A.1
Median trial duration from start of phase I to end of phase III

This figure shows the sum of median trial durations for Phases I, II, and III, for each therapeutic area in the Trialtrove dataset. Each panel corresponds to a different therapeutic area and presents medians for trials taking place in the United States as well as for all trials in the dataset.

Table A.1
Median Trial Duration by Therapeutic Area

This table presents median trial durations (in years) by therapeutic area. Panel A uses the full Trialtrove database, while Panel B uses only trials taking place in the United States. Trial durations are calculated as the difference between the trial start and end dates, divided by 365.

	Phase I	Phase II	Phase III	I + II + III
Panel A: All countries				
Oncology	2.07	2.92	3.61	8.61
Ophthalmology	0.83	1.42	1.59	3.84
Cardiovascular	0.23	1.67	1.92	3.82
CNS	0.34	1.59	1.75	3.67
Metabolic/Endocrinology	0.29	1.49	1.65	3.43
Autoimmune/Inflammation	0.41	1.42	1.56	3.39
Genitourinary	0.32	1.35	1.67	3.34
Vaccines (Infectious Disease)	1.09	0.99	0.84	2.92
Infectious Disease	0.42	1.09	1.27	2.78
All therapeutic areas	0.53	1.99	1.77	4.30
Panel B: United States				
Oncology	2.63	3.01	3.37	9.01
Cardiovascular	0.44	1.72	2.18	4.35
CNS	0.58	1.75	1.89	4.22
Genitourinary	0.49	1.65	1.91	4.05
Ophthalmology	1.06	1.28	1.70	4.04
Autoimmune/Inflammation	0.67	1.63	1.69	3.98
Metabolic/Endocrinology	0.45	1.63	1.84	3.92
Vaccines (Infectious Disease)	1.42	1.25	1.10	3.77
Infectious Disease	0.74	1.40	1.61	3.75
All therapeutic areas	1.20	2.17	2.00	5.37