Distilling Data from Large Language Models:

An Application to Research Productivity Measurement

Maya M. Durvasula[⋆] Sabri Eyuboglu[†] David M. Ritzwoller[‡]

ABSTRACT. We develop a method for assigning high-quality labels to unstructured text. The method is based on fine-tuning an open-source language model with data extracted from a proprietary large language model. We apply this method to construct a census of published clinical trials. We revisit a literature that contends that pharmaceutical research productivity is declining, based on measured increases in the quantity of clinical trials, which outpace trends in output. In our data, the quantity and composition of trials are stable since 2010. Previous measurements are an artifact of biases driven by shifts in the composition of other forms of research.

Keywords: Large Language Models, Medical Research Productivity, Clinical Trials
JEL: C81, O32

Date: September 11, 2024
⋆ Stanford Department of Economics and Stanford Law School, maya.durvasula@stanford.edu
† Stanford Department of Computer Science, eyuboglu@stanford.edu
‡ Stanford Graduate School of Business, ritzwoll@stanford.edu
A previous version of this paper circulated with the title "Medical Research as a Productivity Indicator." We thank Marcella Alsan, Nicholas Bloom, Agnes Cameron, Jiafeng Chen, Liran Einav, Matthew Gentzkow, Han Hong, Guido Imbens, Ramesh Johari, Charles Jones, Neale Mahoney, Evan Munro, Lisa Larrimore Ouellette, Christopher Ré, Joseph Romano, Brad Ross, Bhaven Sampat, Dean Stratakos, Michael Weisman, Heidi Williams, Brandon Yang, Frank Yang, James Zou, and seminar audiences at the NBER Innovation Information Initiative and the Stanford Department of Medicine for helpful comments and conversations. We are especially grateful to Arjun Desai and Karan Goel for their work on the technical infrastructure that enabled this project. We gratefully acknowledge financial support from the National Science Foundation through grant DGE-1656518 (Durvasula, Eyuboglu, and Ritzwoller), the Knight Hennessy Scholars Program, the Stanford Law School John M. Olin Program in Law and Economics, the National Bureau of Economic Research Innovation Information Initiative Summer Fellows Program, and the OpenAI Researcher Access Program. The data produced in this paper are available at the link: https://github.com/DavidRitzwoller/pubmed_clinical_trials.

> Perhaps the most serious task facing empirical work in the area of "technological change" and "invention and innovation" is the construction and interpretation of measures (indices) of advances in knowledge. — *Pakes and Griliches, 1980*

1. Introduction

Considerable progress in the measurement of innovation has been made in the half-century since the publication of Pakes and Griliches (1980). This progress has been enabled in large part by the proliferation of large-scale corpora of patents and scientific publications. Yet measurement remains an active constraint for empirical research (Bryan and Williams, 2021). Patents and publications are attractive proxies for innovation, as they necessarily relate to instances of invention and research. Their utility is limited, however, by heterogeneity in their technical content and value (see e.g., Pakes, 1986; Griliches, 1990, 1994). Consequently, trends in counts of documents are often confounded by trends in their composition (Lerner and Seru, 2022). Correction of this confounding can be viewed as a classification problem: how accurately can unstructured text be mapped to specific quantities of interest?[1]

In this paper, we demonstrate that large language models can facilitate classification tasks of this kind, achieving an accuracy and precision comparable to hand-made labels at a research-relevant scale and budget. We develop a procedure for constructing task-specific large language models, particularly suited to data construction problems that involve classification. The class of methods that we consider has two steps. First, a comparatively small number of labels are extracted from a large, typically proprietary, language model.[2] Second, these labels are used as training data for a smaller, more efficient, open-source language model.[3] This process is an instance of model distillation, which is a growing area of methodological research in computer science (Taori et al., 2023; Xu et al., 2024). This procedure allows researchers to capture the benefits of complex, frontier models, while retaining the resource efficiency and transparency of lighter-weight open-source models.

The primary contribution of this paper is the application of this procedure in a context that is emblematic of the econometric challenges faced by studies of innovation: measurement of trends in clinical trial production. We show that standard methods for data classification identify clinical trials in publication data imprecisely.

[1]The most prominent example of such a classification effort is the construction of the National Bureau of Economic Research (NBER) Patent Data, which, among other things, maps patent records to technology classes (Hall et al., 2001). Earlier, Scherer (1984) classified, by hand, over 15,000 patents according to their technological category.

[2]By proprietary large language model, we are referring, at present, to models such as OpenAI's GPT 3.5 and GPT-4. Alternative models include Anthropic's Claude 3.5 Opus, Google's Gemini, among many others.

[3]Here, we are referring to classes of open-source language models such as RoBERTa, based on Google's BERT model, among others.

These imprecise measurements indicate, spuriously, that the quantity of clinical trials has increased sharply since 2010. In our data, the quantity, quality, and composition of clinical trials have been stable year-on-year.

When considered alongside recent evidence on the stability of drug development costs during this period (Sertkaya et al., 2024), these new measurements of clinical trial production call into question the conclusions of an influential literature that finds diminishing returns to pharmaceutical research (Scannell et al., 2012; DiMasi et al., 2016; Bloom et al., 2020).[4] The conclusions of this literature are based on measures that suggest that the quantity of clinical trials is increasing substantially, in a way unmatched by trends in measures of outputs. In light of our new data, we find that there is little rigorous empirical support for falling pharmaceutical research productivity in the most recent period. This finding is directly relevant to a set of active policy debates, which were initiated, in no small part, in response to concerns that the production of innovative output in this sector is slowing down. These debates include proposals that would restructure public subsidies administered by the National Institutes for Health (NIH) (e.g., Executive Office of the President, 2019; Moss et al., 2020; The Economist, 2022), relax the regulatory standards for approval of new medicines by the U.S. Food and Drug Administration (FDA) (e.g., Roy, 2021; Chertman, 2023; Graboyes and Reimschisel, 2017), and prevent the Centers for Medicare and Medicaid Services (CMS) from negotiating drug prices (e.g., York, 2023; Long, 2023; Inflation Reduction Act, 2022).

Clinical trials are an especially attractive measure of research in medicine (Bryan and Williams, 2021). In the pharmaceutical industry, clinical trials are discrete, regulated experiments that generate useful statistical information about new and existing medicines (Chavez-MacGregor and Giordano, 2016). Trials top the hierarchy of medical evidence employed by firms, regulators, physicians, and patients (Jones and Podolsky, 2015). Since the early 2000s, clinical trials have been reported in a consistent, standardized format across scientific publications and, with less consistency, in administrative databases (DeVito et al., 2020; DeVito and Goldacre, 2021). Trials have, thus, been widely used to measure relative and absolute differences in the composition of research investment (see Finkelstein (2004) and Budish et al. (2015) for two influential examples). Studies of information production and disclosure have examined the substantive content of trials (see e.g., Alsan et al. (2024), Oostrom (2024), and Kao and Oostrom (2024)). Recently, and prominently, measures of clinical trial quantities have been used to support arguments that medical and pharmaceutical research productivity are declining over time (Scannell et al., 2012; Bloom et al., 2020).

Nevertheless, it is remarkably difficult to generate accurate measures of trends in clinical trial production. Administrative databases—such as ClinicalTrials.gov—have experienced large shifts in compliance with

[4] Goldin et al. (2024) survey evidence related to declining productivity across the economy and note that, in the pharmaceutical sector, evidence for declining productivity primarily relies on measures of total numbers of clinical trials and estimates of R&D expenditures.

reporting regulations over time, which make it difficult to distinguish genuine increases in research effort from greater adherence to reporting requirements (see e.g., DeVito et al., 2020). Proprietary databases (e.g., Cortellis, 2024) take as an input the contents of publicly-maintained registries and, thus, are confounded by the same shifts in compliance.

Given these limitations, scientific publication data have served as a natural alternative or complement in recent empirical studies (e.g., Lichtenberg, 2018; Bloom et al., 2020; Kao, 2024; Guzman et al., 2024). With unstructured publication text data, the challenge becomes, again, one of classification: of the millions of records in large-scale publication databases, which disclose the results of a clinical trial? Manually identifying clinical trial records at a scale sufficient to study aggregate trends is infeasible.[5] We show that standard methods—based on publication metadata and machine learning tools—perform poorly. In particular, we show that keyword-based searches, including methods employed by Bloom et al. (2020), are very imprecise and generate spurious measurements of increases in the quantity of clinical trials.[6] Specialized classifiers developed for use in the health services literature, as in Thomas et al. (2021), similarly lack the precision necessary to recover meaningful trends.

Instead, we leverage recent advances in generative AI, which enable data extraction from unstructured text at scale. Our objective is to identify all records in the National Library of Medicine's PubMed / MEDLINE database, from 2010 forward, that disclose the results of a clinical trial that studied the effects of a medicine in human subjects. We aim to classify records based only on the contents of the publication's abstract, with an accuracy close to that of a human labeler.[7]

We proceed in stages. Given the abstract of a scientific publication, we prompt two proprietary language models, OpenAI's GPT-3.5 and GPT-4 (Ouyang et al., 2022; Achiam et al., 2023; Bubeck et al., 2023), to label whether the record satisfies our desired sample restrictions. We find considerable variation in performance with seemingly small differences in prompt text. We iteratively revise prompt text until small changes yield no improvement in performance. GPT-4 performs quite well off-the-shelf. The performance of GPT-3.5 is unsuitable for our application. However, use of GPT-4 at scale—to classify millions of

[5]A small number of papers curate select samples of clinical trial data, for specific research questions. For example, Oostrom (2024) collects records of psychiatric clinical trials as identified in prominent meta-analyses for a study of the relationship between sources of clinical trial funding and reported drug efficacy.

[6]Specifically, Bloom et al. (2020) categorize many records as clinical trials that discuss trials (e.g., meta-analyses and literature reviews). We find that changes in the scientific publishing landscape, discussed in Section 4, increase the size of this categorization error over time.

[7]Many records in PubMed / MEDLINE have full-texts available for download, via the PubMed Central database. However, these full-texts are subject to copyright restrictions. Roughly one quarter of records in PubMed—8 million of the 34 million total records—have full-texts that are available under licenses that would permit this type of research use (the PubMed Central Open Access Subset, at `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`). Thus, we restrict our focus to abstracts, for which no such restrictions apply.

records—is computationally and financially impractical. We classify a moderate number of abstracts with the best-performing prompts. The labels extracted from the proprietary language models are used as training data to fine-tune a set of open-source models. Our preferred sample is constructed using an ensemble of fine-tuned models, which matches the performance of the best proprietary model. False positive and false negative rates are below 5 percent. When this model is used to label the full sample of PubMed records, it yields a sample of approximately 150,000 publications that report the results of clinical trials from 2010 to 2022.

Three technical features of this fine-tuning process are worth highlighting. First, we document a phase-transition in model quality in the number of training labels used. Put differently, we observe a threshold in the number of training labels used in fine-tuning, beyond which the performance of the fine-tuned model increases dramatically. Second, models fine-tuned with labels extracted from GPT 3.5 exceed the performance of GPT 3.5. Third, models fine-tuned with labels extracted from GPT-4 match the performance of GPT-4 itself. That is, we show that researchers can distill the full benefits of large proprietary models into light-weight, resource-efficient models that can be realistically applied to standard data construction tasks.

The resulting data allow us to document two sets of facts about the quantity of medical research over time. First, we find that the quantity of clinical trials has been effectively constant, year-on-year, across our period of interest. We collect a set of ex-ante and ex-post quality measures, derived from scientific publication data. On these dimensions, we find considerable heterogeneity that is, to our knowledge, new to the literature. Roughly half of clinical trial publications in our sample are never cited by a leading medical journal. Nearly 15 percent are never cited by any other scientific publication. However, we find that the distribution across clinical trials of measures of quality, importance, and geography is unchanging.

Second, our data allow us to shed light on the source of the classification errors that drive the trends documented in Bloom et al. (2020).[8] As similar trends in the total quantity of scientific publications have been documented elsewhere, including in Park et al. (2023) and Chu and Evans (2021), these insights may be more generally applicable to discussions about how large increases in the quantity of publications should be interpreted. We contrast trends in clinical trial production with trends in the production of medical research that is similar—conceptually and in its textual presentation in scientific publications—to clinical trials. In our preferred version, this set of other medical research is defined as the set of papers that *cite* clinical trials.[9]

[8]These data are relevant outside of this particular context, for studies of medical innovation and for studies in the health services literature that use clinical trial publications to construct prescribing guidelines for physician practice. Identifying trial records in publication data has been a challenge for this health services literature, where existing best practices have both high true and false positive rates. See Thomas et al. (2021).

[9]In Section 4, we inspect the contents of this set and document that nearly all records fall into one of three categories: observational studies, preclinical studies, and meta-analyses/literature reviews.

While the quantity of clinical trials is constant over this period, the quantity of this other non-trial research has increased by a factor of two.

Three facts about shifts in the composition and quality of these publications provide insight into the source of measured increases in research quantity. First, the large increase in the total number of non-trial papers in our sample is largely driven by researchers in China. Second, this rise in quantity coincides with a decrease (on available measures) in the average quality of publications. Third, there is a substantial increase in the quantity of meta-analyses and literature reviews. The main conclusion that emerges from this analysis is that errors in the classification of scientific publications can yield spurious findings about trends in research effort.

1.1 Related Literature

We draw on, and contribute to, a long literature on the measurement of invention. Decades of work have focused, in particular, on the development of proxies for the *outputs* of technological innovation. An especially prominent approach, due to Hall et al. (2000), uses citation-weighted counts of patents. Our focus, by contrast, is on the difficulties associated with measuring the *inputs* of innovation, which can similarly bias estimates of macroeconomic trends.

The new methods developed in this paper yield measures that call into question a fact at the center of debates involving health, innovation, and fiscal policy. Discussions of declining medical and pharmaceutical research productivity have been especially extensive in the popular press. Collison and Nielsen (2018), Broad (2023), Thompson (2021), and Piper (2023), for example, interpret trends in pharmaceutical research productivity, alongside patterns collected from other fields, as a "warning sign" for the future of technological progress (Piper, 2023).

The idea that patterns like those considered in this paper—rising research quantities and stagnant outcomes—signal the "exhaustion of inventive opportunities" (Griliches, 1994) is old and closely tied to the development of methods for the measurement of innovation. Early appearances of this concern in the literature include Robert K. Merton's 1935 discussion of the "possibility of a slackening in the rate of technologic advance" (Merton, 1935), Alfred B. Stafford's 1952 paper titled "Is the Rate of Invention Declining?" (Stafford, 1952) and Jacob Schmookler's various investigations (Schmookler, 1952, 1954, 1966) into patterns suggestive of diminishing returns to research. Our treatment of these patterns hews closely to that of earlier literatures, which view the evidence base on productivity decline as inconclusive, given the shortcomings of available data (e.g., Pakes and Griliches, 1980; Griliches, 1990, 1994; Cockburn, 2006).[10]

[10] Griliches (1990) summarizes his version of the conclusion that any finding of diminishing returns to innovative efforts are limited by the availability of high-quality data more prosaically: "One can always worry that the world is coming to an end. Someday it undoubtedly will, but it does not look as if the end is already upon us, at least not yet."

This paper differs on one important dimension. We show that advances in data construction technology, if deployed carefully, can allow researchers to resolve some long-standing data construction challenges.

A nascent literature investigates the use of large language models in economics (Dell, 2024; Korinek, 2023). We join a small set of papers—notably, Bartik et al. (2023) and Dell (2024)—that use generative AI as a tool to construct high-quality economic data. Our approach is distinct, in that we fine-tune an open-source language model using data extracted from a larger, proprietary model. In doing so, we show that, with model distillation, it is feasible for researchers to achieve the quality of hand-constructed data at a practically interesting scale, with a reproducible and transparent methodology, at a feasible cost. Prominent examples of model distillation in the computer science literature focus on the construction of chat bots by fine tuning open source models with data extracted from proprietary models (Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023). Other papers have fine-tuned models for simple completion tasks like arithmetic (Liu and Low, 2023) or querying an API (Patil et al., 2023). We are distinguished from this literature by considering a problem with, arguably, a higher-degree of complexity: the construction of data of sufficient quality to be of substantive interest to social science.

Our use of large language models is also distinct from other applications considered in papers such as Korinek (2023), Manning et al. (2024), Bybee (2023), Horton (2023), and Mullainathan and Rambachan (2024), which focus on the use of generative AI for other aspects of the scientific process, including hypothesis identification, survey participation, idea generation, and scientific writing. It is worth emphasizing that, in this paper, we develop a procedure that is applicable to a binary classification problem, albeit one that is highly complex. The viability of these methods, and the extent of required modifications from the process outlined here, for application in more complicated classification problems of interest to economists, including those outlined in Dell (2024), is a valuable direction for further research.

More generally, we contribute to the large empirical literature that uses text as data (Gentzkow et al., 2019). We anticipate that our procedure for the construction of new, precise data on scientific research may be applicable to other contexts where unstructured text is available and manual construction of data is infeasible. We highlight a set of steps—in validation, prompt design, and fine-tuning of open source models—that yield high-quality data at comparatively low cost.

2. Identifying Clinical Trials

We construct a new census of clinical trials disclosed in scientific publications. We begin, in Section 2.1 and Section 2.2, by providing context on the objectives of our data construction exercise and by detailing the deficiencies of existing, or alternative, approaches to measuring this form of medical innovation. In

Sections 2.3 and 2.4 we propose a method, based on distilling data from a large language model, for identifying relevant records at scale and with high fidelity. The performance of the method is assessed in Section 2.5.

2.1 Context

Brief institutional context on the existing evidence related to pharmaceutical research productivity is helpful both in motivating our data construction and in interpreting our findings.

Declining pharmaceutical research productivity has been a topic of policy concern since the late 1990s. In the past thirty years, discussions have largely centered on two sets of empirical patterns, which Cockburn (2006) summarizes as follows: "More and more money is being invested in R&D, but the rate at which new drugs are introduced is failing to keep pace." R&D, in this setting, refers to a particular type of biomedical research: investments in preclinical and, especially, clinical studies intended to translate discoveries made in basic science into commercialized medicines suitable for use in human patients. That inputs per new drug appear to be rising sharply over time is viewed, generally, as an established fact in the empirical productivity literature (Bloom et al., 2020; Goldin et al., 2024).

Measures of productivity in this setting are somewhat crude. The measure of outputs in the most influential articulation of this phenomenon, by Scannell et al. (2012), is the number of "new molecular entities"—innovative drugs—approved by the U.S. Food and Drug Administration.[11] As Bloom et al. (2020) note, that this count has been stagnant or declining is viewed as an established fact.[12]

Measures of inputs are, typically, some estimate of R&D expenditures. The simplest construction of a measure of R&D expenditure scales the (inflation-adjusted) cost of a clinical trial by the number of clinical trials. Recent evidence on these objects comes from Bloom et al. (2020) and Sertkaya et al. (2024). Bloom et al. (2020) find a sharp increase in the number of clinical trials, as measured in publication data over time. Using granular data on clinical trial site contracts and internal records of the U.S. Food and Drug Administration, Sertkaya et al. (2024) find that the cost of clinical trials and the cost of drug development was roughly constant between 2000 and 2018.

Arguments of declining productivity in the pharmaceutical industry, then, depend substantially on the finding that the quantity of clinical trials is rising.

[11]Technically, a new molecular entity is a drug that contains no previously approved active moiety. In practice, this means that every active element of the drug (e.g., acetaminophen and caffeine in the commonly used over-the-counter migraine medicine Excedrin) must be novel.

[12]In data on new molecular entities approved by the FDA between 2005 and 2021, we observe a small increase in the number of approvals, largely driven by an uptick in approvals in 2020 and 2021. In earlier periods of data, we confirm that there is no discernible trend in new molecular entity approvals.

2.2 Existing Approaches

Although many datasets report information on clinical trials—and have been increasingly used by economists to measure innovation (see e.g., Budish et al., 2015; Cunningham et al., 2021)—no existing dataset, used off-the-shelf, yields accurate, transparent aggregate quantity measures.

2.2.1 Trial Registries and Proprietary Databases. Nearly all clinical trials are conducted under the auspices of a regulator that requires both pre-registration and results disclosure. In the United States, the Food and Drug Administration Amendments Act (FDAAA), as passed in 2007 and revised in a "Final Rule" that took effect in 2017, imposes these requirements on sponsors of regulated trials. Nominally, effected sponsors are required to register their studies on "ClinicalTrials.gov." For nearly all studies, sponsors must also report the results of a trial within a designated period after completion. Moreover, increasingly, medical journals require registration as a pre-condition of publication (Laine et al., 2007).

In practice, there is widespread noncompliance with these regulations and requirements. Deborah Zarin, then-director of ClinicalTrials.gov, notes that "funders, sponsors, and institutional review boards continue to allow unregistered trials or trials with late registration to be conducted, and some journals continue to allow the results of such trials to potentially be published" (Zarin et al., 2017). Noncompliance with federal reporting requirements is similarly widespread. In a series of papers assessing compliance over time DeVito et al. (2020) and DeVito and Goldacre (2021) find that roughly 60 percent of trials registered on ClinicalTrials.gov fail to report required data elements at the time of their reporting deadline.[13]

From the perspective of our study, there are three issues with the use of registry data. First, requirements for study registration have changed over time. Table B.1 lists five changes between 2000 and 2020 that altered the set of studies that should appear in a complete version of the database. Second, sponsors engage in considerable "back-filling" of past studies. Zarin et al. (2017) find that a large proportion of trials are registered after the trial start date.[14] Third, not all records in ClinicalTrials.gov and similar registries are, in fact, clinical trials (Tse et al., 2018). As a result, it is difficult to determine whether changes in the number of records reflect actual growth in scientific research or a change in the propensity to register studies (trials and non-trial studies).

[13]Note that noncompliance in this industry persists despite the existence of high statutory penalties. Sponsors can be fined up to $10,000 per day for failing to register studies and disclose results. See the FDA Amendments Act of 2007 (FDAAA) and its Final Rule issued in 2021, at 42 U.S.C. § 282(j)(5)(C)(i). To date, no fines have been collected, but the FDA now sends notices of noncompliance to delinquent sponsors (Stephenson, 2021). Notices are available here: https://www.fda.gov/science-research/fdas-role-clinicaltrialsgov-information/clinicaltrials gov-notices-noncompliance-and-civil-money-penalty-actions

[14]Often back-filling relates to changes in organizational disclosure policy. In a version of ClinicalTrials.gov current through 31 May 2023, Boehringer Ingelheim—the largest privately held pharmaceutical company in the world—appears to have posted 804 studies in 2014, only 72 of which were started in 2014. In 2013, Boehringer registered 101 studies. In the fifteen preceding years, since the creation of the database, Boehringer posted 872 records in total. We selected this case study based on a discussion in Zarin et al. (2017).

In Appendix B, we illustrate the difficulty of recovering accurate measures of trends in the quantity of clinical trials from this database. We show that small changes in data construction yield meaningfully different estimates of trial levels and that many records are not in fact clinical trials. Moreover, due to widespread data gaps resulting from noncompliance in reporting of results and study attributes, there is little reliable information about characteristics of registered studies.

Other databases, such as the PDQ Cancer Information database maintained by the U.S. National Cancer Institute and used in papers including Budish et al. (2015), are disease-specific and, thus, insufficiently general for our purposes. A small number of firms (notably, GlaxoSmithKline) have disclosed all clinical trials over a longer period of time.[15] Of course, records from a single firm are not informative about aggregate trends, especially because many investigational drugs change hands frequently during the course of development (Cunningham et al., 2021).

Researchers, instead, often draw on proprietary databases. Although these databases are sufficient for certain research questions—for example, related to changes in the ownership of investigational drugs (Cunningham et al., 2021, using Pharmaprojects data) or on the link between insurance policy and drug development (Agha et al., 2022, using Cortellis data)—they are unsuitable for our particular productivity measurement exercise. Our aim is to construct a census that allows for comparison of quantities across points in time. First, we are unable to determine, with specificity, how records are identified and catalogued based on available documentation. Second, these databases take public registries as inputs and, thus, risk adopting the same shifts in reporting over time (see e.g., the list of registry inputs for one frequently-used database, Cortellis, 2024).

2.2.2 Publication Data. Publication data provide a natural alternative.[16] The database PubMed / MEDLINE indexes roughly 34 million publications, a near universe of published biomedical research.[17] The challenge is to classify the specific forms of research captured by this corpus. Of these 34 million records, which are clinical trials?

The existing literature takes several approaches. In health sciences—where data on the universe of clinical trials are relevant for the construction of meta-analyses and the development of prescribing guidelines—researchers typically employ over-inclusive methods with high true positive rates. To use one prominent example, Cochrane, a British organization that synthesizes the findings of medical research, developed the

[15] See GlaxoSmithKline's registry here: https://www.gsk-studyregister.com/

[16] In Section 4, we discuss the two natural concerns with measures constructed from publication data—publication bias and shifts in reporting norms over time. Neither is likely to affect the interpretation of trends in our data.

[17] See Appendix A.1 for further details on on our treatment of the PubMed data.

"Cochrane RCT Classifier," a machine learning-based classifier for retrieving randomized controlled trials. The Cochrane approach successfully identifies trials with a 99 percent true positive rate. For every eight true positives, however, it returns 92 false positives. See Thomas et al. (2021) for details.

Approaches used by researchers in social science, instead, largely draw on metadata contained in PubMed.[18] Lichtenberg (2018) and Bloom et al. (2020) collect all records assigned a "publication type"—in the PubMed indexing process—of "clinical trial." Feldman et al. (2019) collect all records with a publication type similar to clinical trial (e.g., also including "randomized controlled trial"). On inspection, these tags are imprecise. Both approaches retrieve large numbers of records that are not, in fact, clinical trials, while excluding potentially relevant records.

Figure 2.1 contrasts counts of published trials constructed in various ways. The light green line ('Clinical Trial' NLM Tag) plots the number of publications, in each calendar year, indexed with the publication type "clinical trial." This light green line is the measure reported in Bloom et al. (2020). The teal line (Any NLM Tag) plots counts of records indexed with any of 18 types that are likely to include clinical trials or related medical research.[19] Alternative approaches to identifying trials are similarly imprecise. Medical journals increasingly require trials to be registered in open registries, e.g., ClinicalTrials.gov, as a condition of publication. The blue line (Any Registry Tag) counts records that report, in their abstract, an identifier associated with one of the four largest international trial registries. Language in a record's abstract can indicate that it might be a clinical trial if, for example, it references a "treatment group" and "control group." The dark blue line (Any Keyword) plots the count of records with such keywords over time.

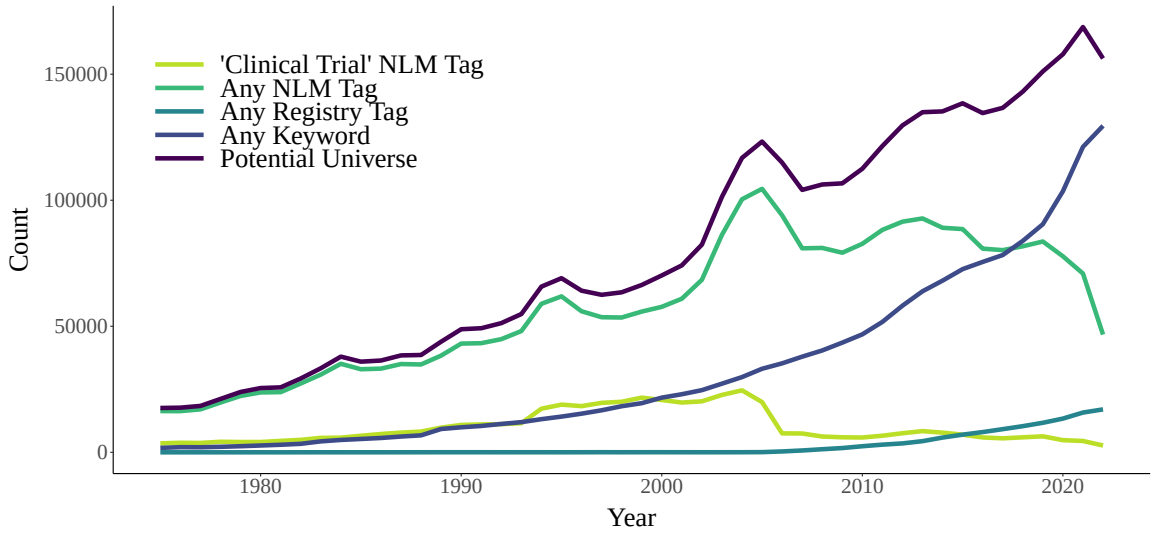
We define the "potential universe" of clinical trials as those publications with any of the following: a publication type variable similar to clinical trial, a registry identifier reported in its abstract, or a keyword reported in its abstract. The purple line (Potential Universe) plots this trend over time.

Figure 2.1 highlights that alternative approaches to data construction, in this setting, yield meaningfully different conclusions about levels, trends, and composition. See Appendix A.2 for further details on the construction of these series. Given these difficulties, we describe a procedure, below, to recover a sample of clinical trials from this corpus of text in a manner that is consistent and transparent.

[18] Oostrom (2024) and Kao et al. (2023) are two exceptions. Oostrom (2024) uses data from a meta-analysis of published clinical trials. Kao et al. (2023) collect trial records from the proceedings of scientific conferences.

[19] On inspection, it appears as though the NLM shifted from primarily using the tag "Clinical Trial" to "Randomized Controlled Trial" and added more specific tags (e.g., "Clinical Trial, Phase 1"). There are 19,937 records tagged with "Clinical Trial" in 2005 and 7,518 with this same tag in 2006.

FIGURE 2.1. Universe of Potential Clinical Trials



Notes: Figure 2.1 displays counts of the number of clinical trials indexed in PubMed / MEDLINE over time, constructed using alternative search strategies. The National Library of Medicine (NLM) categorizes each publication into a "pubtype." The light green line ('Clinical Trial' NLM Tag) displays the number of publications in the "Clinical Trial" pubtype. The teal line (Any NLM Tag) displays the number of publication whose pubtype is an element of a set of 18 categories likely to include clinical trials. The blue line (Any Registry Tag) gives the number of publications that report, in their abstract, an identifier associated with one of the four largest international trial registries. The dark blue line (Any Keyword) indicates the number of publications whose abstract contains a keyword indicative of a clinical trial. The purple line (Potential Universe) displays the number of clinical trials in the union of the sets of publications identified with the other lines. See Appendix A.2 for further details.

2.3 Objective

What exactly are we trying to recover when we search for clinical trials? We argue that the object of interest is a census of clinical trials that study the effects of a medicine in human subjects.

Definition 2.1. The sample of interest is composed of all publications that report the results of a prospective, interventional clinical trial that evaluates the effects of investigational or approved drugs in a setting with exclusively human subjects.

Definition 2.1 embeds several restrictions, intended to yield precise measures of clinical trials relevant to the approval of new pharmaceuticals for use in human patients. Studies involving animals, literature summaries, and re-analyses of existing data, for example, are excluded. We take the "potential universe" plotted in Figure 2.1 as our baseline sample of potential clinical trials.

We further restrict attention to the 1,821,429 records published in or after 2010, as there is a trade-off between data coverage and quality in this setting. In the 1990s and 2000s, medical experts created a standardized system for clinical trial reporting in scientific publications. By the time of the most recent

Consolidated Standards of Reporting Trials (CONSORT) in 2010, nearly all medical journals required authors to comply with a standardized structure for all clinical trial publications (Schulz et al., 2010). As a result, most trial publications have a distinctive format and include a pre-defined set of elements. Although our procedure can be applied to any type of scientific publication text, the standardization required by the CONSORT Statement changes the relationship between clinical trial research and publications in the early 2000s. To answer the research question at the center of our empirical analysis—is the quantity of clinical trials rising over time?—we elect to proceed with a narrower sample to avoid this issue of interpretability.

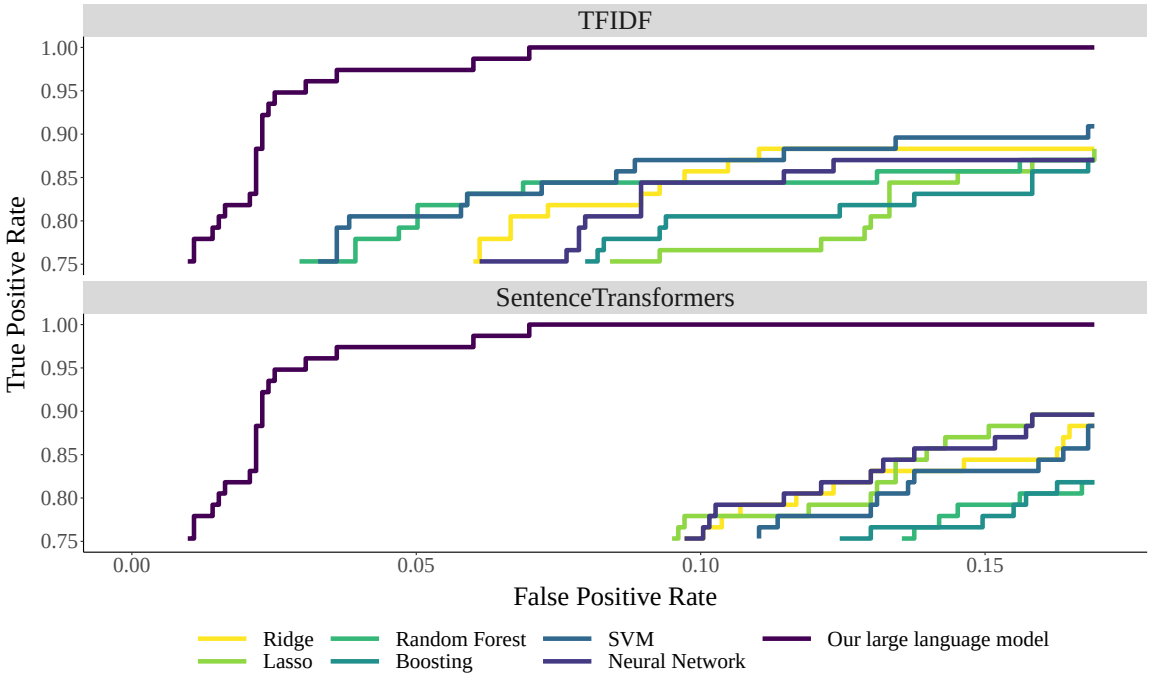
To quantify our ability to identify the clinical trials in this set of publications, we hand-label approximately 3,000 randomly selected publications based on the content of their abstracts. We develop a custom labeling interface that reduces the time required to label 100 records, for the authors, by a factor of seven. Appendix C.1 describes this tool and its application to this context in detail. Of the hand-labeled publications, 11.2% meet the criteria for inclusion in our sample. The hand-labeled data are split into three subsets—validation, training, and testing—based on their eventual use.

As a baseline, we assess the performance of several standard machine learning algorithms. The results are displayed in Figure 2.2. Further details on the construction of this figure are given in Appendix C.2. Each algorithm is estimated on the hand-labels assigned to the publications in the training and validation sample splits. Performance is measured in the testing sample split. For feature vectors, we use either TF-IDF embeddings computed in the corpus of abstracts in the hand-labeled sample or the embeddings of each abstract obtained from the SENTENCETRANSFORMER language model (Reimers and Gurevych, 2019). We find that the performance of the standard machine learning algorithms is again unsuitable for our application. At a 90% true positive rate, the best performing model identifies 50 true positives for every 50 false positives.

2.4 Model Distillation

We construct a large language model optimized for our task. This is accomplished in three steps. First, we iteratively construct a set of prompts that exhibit good performance when posed to proprietary models—OpenAI's GPT-3.5 and GPT-4. *A priori*, we expect GPT-4 to output labels of slightly lower quality than hand-labeled output and GPT-3.5 to produce slightly noisier labels than GPT-4. The extraction of GPT-3.5 labels is substantially faster and cheaper. Second, we extract noisy labels for a moderate number of publications in our sample by querying the proprietary models. Third, the noisy labels are used to train an off-the-shelf large language model. The resulting model is then used to identify clinical trials in our baseline sample. This process for building specialized large language models is referred to as model distillation (Xu et al., 2024).

FIGURE 2.2. Performance of Standard Machine Learning Methods



Notes: Figure 2.2 displays the receiver operating characteristic of several machine learning models, trained on two types of embeddings, for classifying whether a publication satisfies the restrictions enumerated in Definition 2.1. The hyper-parameters of each model are determined with 10-fold cross-validation. The models are trained in the training and validation data set. Error rates are estimated in the testing data set. In addition, we display the performance of our, ensemble, fine-tuned large language model in purple. This is the same curve labelled "Ensemble" in Figure 2.4.

2.4.1 Prompt Design. To produce these labels at scale, each of the two models must be appropriately prompted. That is, each pre-trained model must be provided with a block of text as an input and asked to return the appropriate completion of this input. *A priori*, it is not clear what prompt structure will work well. The relative infancy of this area of research renders it difficult to identify a set of "best practices."[20] All of the prompts that we consider are displayed as figures in Appendix E.

We identify three general prompt formats, which differ both in the amount of detail provided about our classification task and in the structure of the requested model completion. The simplest prompt provides a version of our sample definition, Definition 2.1, and the text of an abstract, and asks the model to return 'TRUE' if the abstract satisfies these criteria. Otherwise, it returns 'FALSE'. The text of this prompt is displayed in Figure E.1. We refer to this prompt as Prompt 1.0. A second, more complicated prompt provides the same definition, with a set of examples of publication characteristics that do and do not satisfy the definition. Here, the prompt asks the model to return either 'TRUE' or the name of a specific excluded

[20]See, for one example of this evolving area of work, a guide to "prompt engineering" from OpenAI: https://platform.openai.com/docs/guides/prompt-engineering.

category. The text of this prompt is displayed in Figure E.5. We refer to this prompt as Prompt 2.0. A third, more complex prompt provides the same definitions and examples, but asks the model to return either 'TRUE' or an explanation of why the record does or does not satisfy our sample definition. The text of this prompt is displayed in Figure E.7. We refer to this prompt as Prompt 3.0. We devise initial language for each prompt iteratively, using a small number of records from our hand-labelled validation data.

We test each of these three prompts in our 1000-record hand-labelled validation dataset, using both GPT-3.5 and GPT-4. We conduct a detailed error analysis, reported in Table C.1. For each instance in which a model returns a label (TRUE/FALSE) that differs from that in the hand-labelled dataset, we inspect the record. We categorize errors into types and sub-types. Details are given in Appendix C.3.

This exercise is instructive on three margins. First, it highlights the highest performing prompts. Second, it indicates particular types of errors in categorization—which suggest opportunities for more precise language in a prompt. Third, it draws attention to differences in the performance of the two models. Observe that, in the "Other" error type, we include a sub-type called "overly literal interpretation of inclusion criteria." We primary record errors of this type for Prompt 3, which asked the model to return a completion that described why, or why not, a record was classified as being in our sample. Here, GPT-3.5 makes 30 such errors—classifying records as FALSE by recapitulating the sample definition provided. GPT-4 makes two errors. The insight that emerges from this exercise, then, is that different prompt structures may be preferable depending on the model used, and that, by and large, GPT-4 is better suited for tasks that require (the resemblance of) more sophisticated reasoning.

We revise each class of prompt based on these findings. For Prompt 1.0, the simplest true/false prompt, we consider three variants. We refer to these revised prompts as Prompts 1.1, 1.2, and 1.3. The text of these prompts is displayed in Figures E.2 to E.4. For Prompts 2 and 3, we consider one variant each. We refer to these revised prompts as Prompts 2.1 and 3.1. The text of these prompts is displayed in Figures E.6 and E.8. These changes reflect the differences in performance catalogued in Table C.1.

Table 2.1 reports estimates of the true positive rate and false positive rate for each prompt, in both models, computed in the validation sample. Observe—for example, with Prompts 1.1, 1.2, and 1.3 queried to GPT-4—that even small changes in the text of a prompt yield substantial differences in performance. We attempt a second round of prompt iteration (Prompt 1.3). We observe that performance deteriorates with even small modification. Thus, we select the two highest performing prompts—one for each proprietary model. We use Prompt 2.0 to extract weak labels using GPT-3.5, and Prompt 1.2 to extract weak labels using GPT-4.

TABLE 2.1. Prompt Performance in Validation Data

Panel A: GPT-3.5

Prompt Type	Prompt Sub-Type	False Positive Rate	True Positive Rate
1	0	0.247	0.788
	1	0.172	0.898
	2	0.037	0.584
	3	0.037	0.489
2	0	**0.162**	**0.876**
	1	0.176	0.905
3	0	0.248	0.722
	1	0.171	0.883

Panel B: GPT-4

Prompt Type	Prompt Sub-Type	False Positive Rate	True Positive Rate
1	0	0.202	0.971
	1	0.065	0.949
	2	**0.049**	**0.934**
	3	0.056	0.912
2	0	0.081	0.964
	1	0.072	0.964
3	0	0.167	0.971
	1	0.068	0.956

Notes: Table 2.1 records the performance of each of eight prompt variants in a sample of 1,000 validation dataset records. Panel A reports performance associated with the proprietary model GPT-3.5. Panel B reports analogous statistics for model GPT-4. Prompt 1 asks the model to return TRUE or FALSE. Prompt 2 asks the model to return TRUE or the name of a specific, excluded category. Prompt 3 asks the model to return TRUE or an explanation of why the record should be excluded. Prompt sub-types correspond to various iterations. Sub-type 0 is the initial version of the prompt. Sub-types 1-3, where applicable, are subsequent iterations. Appendix E records the text of each prompt. We use Prompt 2, Sub-Type 0, to extract weak labels using GPT-3.5, and Prompt 1, Sub-Type 2 to extract weak labels using GPT-4. The performance measurements for these prompts are displayed in boldface.

The performance of GPT-4 represents a substantial improvement over existing methods. However, practical and substantive considerations mitigate the applicability of proprietary models for classification of the complete baseline sample. Practically, it is—and likely will remain—prohibitively expensive, both computationally and financially, to deploy GPT-4 at this scale.[21] Substantively, proprietary models are black boxes. Their details and substance are not public and are known to change at regular intervals.

2.4.2 Fine-Tuning. We compute noisy labels for a moderate number of randomly selected publications in our baseline sample using the best performing prompts for both GPT-3.5 and GPT-4. These noisy labels are

[21]We incur a cost of roughly $4,500 to extract noisy labels for 64,000 abstracts. Extrapolating this figure to the full sample of 1.8 million abstracts gives a price of approximately $130,000.

used as data to train off-the-shelf BERT models from two architecture classes: (1) BIGBIRD (Zaheer et al., 2021) and (2) BIOMEDBERT (Gu et al., 2021).[22] Both models use medium-scale Transformer architectures, i.e., between 100 and 300 million parameters, pre-trained with the masked-language modeling (MLM) objective (Vaswani et al., 2023; Devlin et al., 2019). The models differ in their architectural details and pre-training corpora.

BIGBIRD uses a sparse attention mechanism to reduce the computational cost of processing long text sequences.[23] This enables efficient handling of long documents up to 4,096 tokens. In contrast, standard BERT models can only process 512 tokens. BIGBIRD is useful because abstracts of clinical trials regularly exceed 512 tokens. Prior to pre-training, the model was warm-started from the ROBERTA checkpoint (Liu et al., 2019) and then pre-trained on the standard BERT corpus.[24] We evaluate two model sizes: BIGBIRD Base (125 million parameters) and BIGBIRD Large (355 million parameters).

BIOMEDBERT uses a domain-specific pre-training corpus sourced from articles on PubMed (Gu et al., 2021). It uses the same model architecture as RoBERTa, but a different, domain-specific token vocabulary optimized for PubMed articles. Like BIGBIRD, we evaluate two model sizes: BIOMEDBERT Base (125 million parameters) and BIOMEDBERT Large (355 million parameters).

We fine-tune the pre-trained architectures to classify abstracts according to the inclusion and exclusion criteria enumerated in Definition 2.1. We replace the language modeling classification head with a binary classification head consisting of a single linear layer and softmax activation. We fine-tune the full model (*i.e.* not just the classification head) with cross-entropy loss.[25] All models were trained on a single $V100$ GPU with 16GB of HBM.

2.5 Performance

Given the text of an abstract, our fine-tuned language models output a probability that the publication satisfies the restrictions enumerated in Definition 2.1. Publications whose probabilities fall above a chosen threshold are classified as belonging in our sample. Figure 2.3 displays estimates of true and false positive rates, computed with the test data, as we vary this threshold for labels assigned by fine-tuned version of the

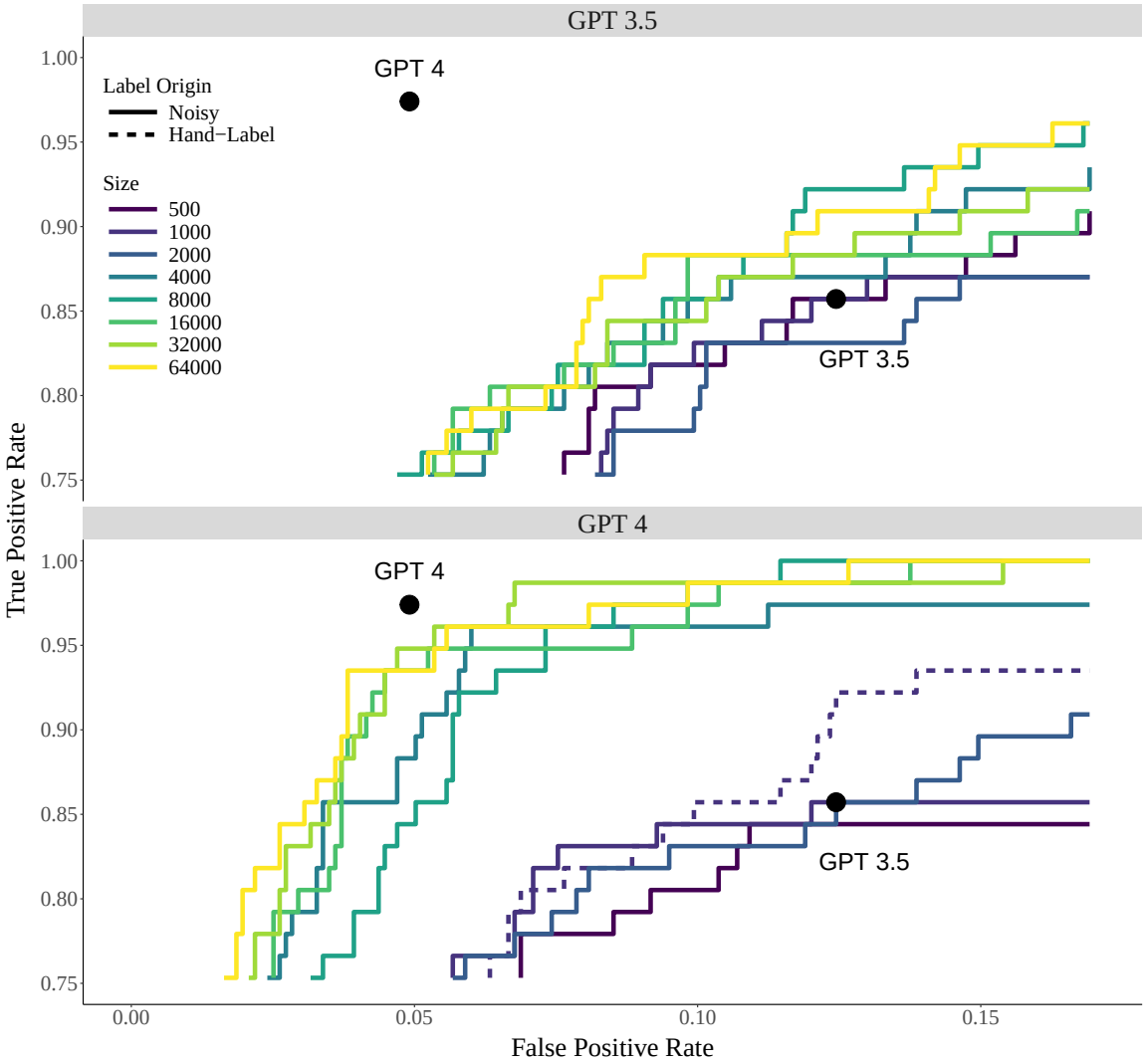
[22]There are many open-source large language models (*e.g.*, LLaMa, Mistral, Pythia). Many are several orders of magnitude (*e.g.*, 7-70 billion parameters) larger than BERT ($\sim$ 350 million parameters). We selected BIGBIRD and BIOMEDBERT via trial-and-error that included testing the performance of these larger models, including LLaMA (both 7 and 70 billion parameter versions, trained with QLoRA). Both models exhibited comparable performance, but BIGBIRD was substantially faster and simpler to train.

[23]The computational complexity of regular attention is quadratic in sequence length, while BIGBIRD reduces this to linear complexity.

[24]The standard BERT corpus consists of the Books (Zhu et al., 2015), CC-News (Guu et al., 2020), Stories (Trinh and Le, 2019), and Wikipedia (Wikimedia Foundation, 2023) datasets.

[25]We use the Adam optimizer with learning rate 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ (determined via hyperparameter sweep). We use a maximum sequence length of 4096 tokens. If an abstract does not fit into the maximum sequence length provided by a model, the abstract is truncated.

FIGURE 2.3. Receiver Operating Characteristic by Data Size and Origin

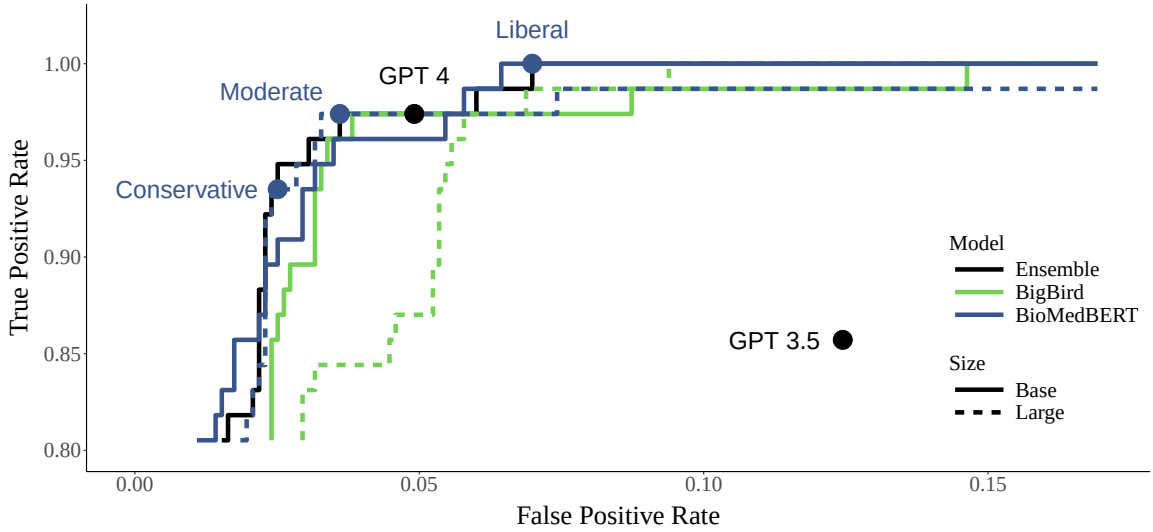


Notes: Figure 2.3 displays estimates of the the receiver operating characteristic of several fine-tuned language models, based on the BIGBIRD architecture. The models differ in terms of the source of their training data. Solid lines display the performance of models trained on noisy labels extracted from either GPT-3.5 or GPT-4. The dotted line displays the performance of a model trained on the hand-labeled data in the training dataset. Estimates of the true positive rate and false positive rate of GPT-3.5 and GPT-4 are indicated with black dots and are measured using the testing data.

base BIGBIRD model. The top and bottom panels are trained on noisy labels extracted from GPT-3.5 and GPT-4, respectively. We vary the quantity of noisy labels used to train the language model. Additionally, in the bottom panel, we display the performance of a fine-tuned model trained with the 1000 hand-labelled observations in the training dataset.

Fine-tuned models trained with noisy labels extracted from GPT-3.5 outperform GPT-3.5. By contrast, fine-tuned models trained with noisy labels extracted from GPT-4 match the performance of GPT-4, and

FIGURE 2.4. Receiver Operating Characteristic by Model Architecture



Notes: Figure 2.4 displays estimates of the receiver operating characteristic of several models used to classify the publications indexed by PubMed / MEDLINE according to whether they satisfy the restrictions enumerated in Definition 2.1. All metrics are computed in the testing split of the hand-labelled publications. Estimates of the true positive rate and false positive rate of OpenAI's proprietary models GPT-3 and GPT-4 are indicated with black dots. The curves give the performance of four fine-tuned, open-source large language models, in addition to an ensemble model. The blue dots indicate the true positive rate and false positive rate of the models used to construct the "Conservative," "Moderate," and "Liberal" samples of clinical trials.

significantly outperform models trained with labels extracted from GPT-3.5 or with hand-labels. There appears to be a threshold where performance dramatically improves at around 8,000 training labels.

Figure 2.4 displays analogous estimates for both sizes of the BIGBIRD and BIOMEDBERT models. Both models are trained with 64,000 training labels extracted with GPT-4. Again, the fine-tuned models are able to match the performance of GPT-4 in the test data. The figure additionally displays the performance of an ensemble model estimated in the training data. This model is obtained by estimating a logistic regression of the hand-labels on the probabilities output by all four models displayed Figure 2.4. The ensemble model is used to produce our final sample.

We choose three thresholds according to the stringency with which they enforce the sample restrictions. The estimated true and false positive rates associated with these points are displayed in Figure 2.4 and labeled as "Conservative," "Moderate," and "Liberal." Our preferred sample is associated with the conservative

threshold.[26] For every 82 true positives, the conservative model identifies 18 false positives.[27] We report results associated with the moderate and liberal thresholds in the appendix as tests of robustness. Our final, conservative, sample consists of 152,027 publications classified as satisfying the sample restrictions.

Each table displayed in Appendix A compares the contents of our sample to the counts of potential clinical trials plotted in Figure 2.1. Although certain elements of PubMed metadata—including the union of all NLM tags—capture many of the records in our final sample, we confirm that they miss many records that we flag as trials and include many records that do not satisfy Definition 2.1. No combination of existing search strategies, then, classifies records with the same accuracy or precision as our final sample.

3. Disentangling Trends in Publication Data

Figure 2.1 indicates that, across measures, the total quantity of potentially-relevant records in PubMed has increased substantially over time, including since 2010.[28] These patterns are consistent with documented increases in the quantity of scientific papers in Bloom et al. (2020) and Park et al. (2023), as well as economy-wide increases in research effort reported in Goldin et al. (2024). In this section, we use our census of clinical trials to decompose this trend. We contrast trends in the production of clinical trials with those documented for other, similar forms of medical research.

There are many potential ways to construct a sample to contrast with data on clinical trial production. Our preferred approach defines this category as the collection of publications that *cite* clinical trials. This sample includes publications that "look like" clinical trials—a fact that we confirm on inspection—and which necessarily engage with the findings of these studies. Thus, in an imprecise classification task, it is precisely this set of records that are likely to be erroneously included. For example, a keyword search of publication abstracts for "clinical trials" may capture studies that summarize the findings of clinical trials, without reporting new information. Based on our inspection of these records, these types of errors are especially likely for publications in which the writing quality is low, for which it may be challenging—even for a human labeler—to determine whether the record reports novel results or summarizes existing findings.

[26] The liberal model may be particularly useful for conducting literature reviews, where a near-perfect true positive rate is needed.

[27] In the test data, the conservative model assigns incorrect labels to 27 of 993 papers. We conduct an error analysis. See Appendix C.4 for further details. In 13 cases, there is a clear error. In 14 cases, however, errors are associated with records that were difficult to categorize for a human labeller. For example, PubMed record 32737793 is flagged as satisfying Definition 2.1 with all three thresholds, but is a literature review. By contrast, PubMed record 27880726 was categorized as satisfying Definition 2.1 twice by a human labeller. It is excluded from both the conservative and moderate model-generated samples. On inspection, the abstract does not explicitly state that the study enrolled human subjects, but hints that it may have been conducted in an animal model. Review of the associated full text confirms that this study did, in fact, enroll only rats.

[28] Here, we focus on patterns in the production of publications that are similar to those that disclose the results of a clinical trial. We find similar patterns in PubMed as a whole, and in alternative cuts of records. Across fields, the total number of scientific papers, the total number of unique authors listed on scientific papers, the total number of citations given and received by scientific papers, the total number of scientific journals, and the total number of pages in scientific journals exhibit similar patterns.

Alternative methods of constructing a measure of "other" research introduce additional, complicating conceptual problems. Approaches that rely on keyword searches, NLM tags, etc., as Section 2 and its appendices make clear, capture different sets of records over time, as a consequence of changes in database structure. Approaches that examine other publications in the same journals as clinical trials will mechanically suggest shifts in total quantities, if the number of pages in journals change over time or if certain journals become online-only.[29]

The primary challenge in using a citation-based measure is one of truncation: constructing measures of papers that cite clinical trials requires determining the window in which such citation must have occurred. To account for concerns about citation delays, we construct various measures, using 2–6 year citation counts. Two-year citation counts, thus, count publications that cite clinical trials published in the preceding two years. As our clinical trial sample begins in the year 2010, these measures are available only beginning in 2012. It is possible, and common, for clinical trials to cite other trials. We drop from our "citing" sample any records that we also flag as reporting the results of a trial.

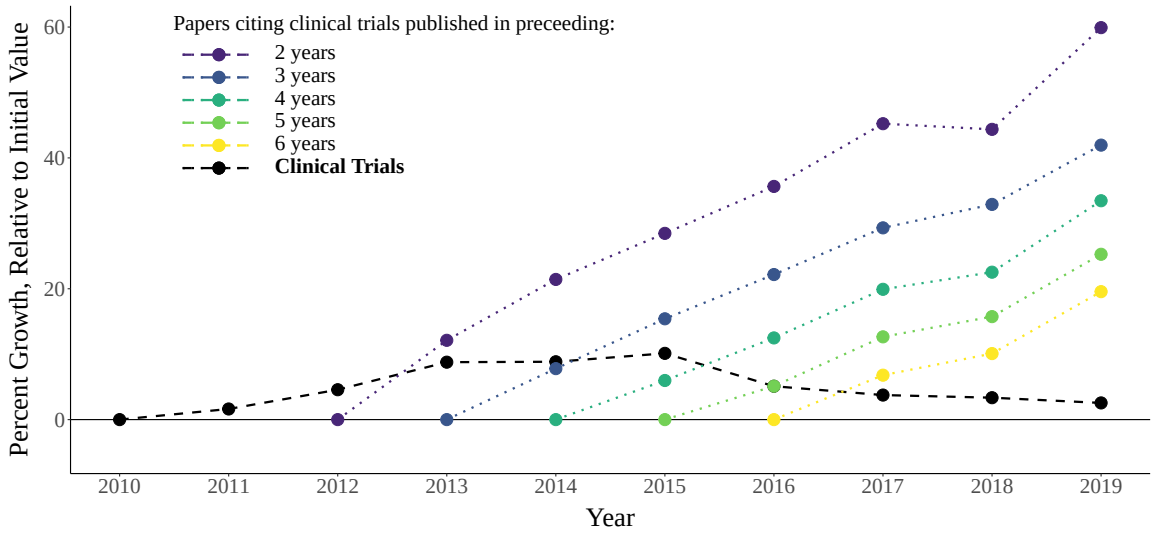
3.1 Decomposing Trends

Figure 3.1 documents that the quantity of clinical trials—the black line—has remained essentially constant over the past ten years. In contrast, the quantity of other non-trial research—the colored lines—has increased substantially. Observe that we display the percent *growth* in the quantity of clinical trials in our sample published in each calendar year, alongside the *growth* in the set of papers that cite clinical trials, constructed in various ways to account for the truncation discussed above. Each series is normalized to begin at zero. This presentation masks substantial differences in the levels of these series. There are 10,903 clinical trials in our census in 2010, and 30,841 publications that cite clinical trials (using three-year citations) in 2013.

These trends are robust to alternative forms of data construction. In Figure D.1, we construct these plots using our liberal and moderate sample definitions and find, qualitatively, no change. In Figure D.2, we explore a potential concern with a citation-based measure. We document in Figure 2.1 that the quantity of scientific publications is rising over time. It is reasonable to expect that this rise, coupled with the ease of citing papers enabled by the internet, coincides with an increase in the total number of citations in the aggregate and per paper. We confirm, empirically, that both hypotheses are borne out by our data. To account for the fact that the "cost" of citing a clinical trial may have decreased over time, thus generating the trend documented in Figure 3.1, we construct a version of this plot in which each citing paper is weighted by the total number

[29] Ioannidis et al. (2023) document that the number of medical journals is increasing. In 2024, the *Annals of Emergency Medicine*, the highest impact journal in the medical sub-field of emergency medicine, shifted to an online-only format. Brainard (2020) discusses dozens of instances in which medical journals transitioned to online-only formats, then ceased publication entirely.

FIGURE 3.1. Growth in Clinical Research, Stability of Clinical Trials



Notes: Figure 3.1 displays measurements of the number of clinical trials, and papers that cite clinical trials, published in each calendar year. Each series is reported in terms of the percent change relative to its initial value. The sample of published clinical trials is constructed with the conservative model. To address truncation, we report the number of publications that cite clinical trials published in the preceding t years for each t between 2 and 6.

of papers that it cites. That is, a citing paper that cites a trial and one other paper receives a weight of 1/2, and a citing paper that cites a trial and nine other papers receives a weight of 1/10. In doing so, we "penalize" papers that are especially generous in their tendency to cite other papers. Figure D.2 suggests, again, that this trend is robust to this alternate construction.

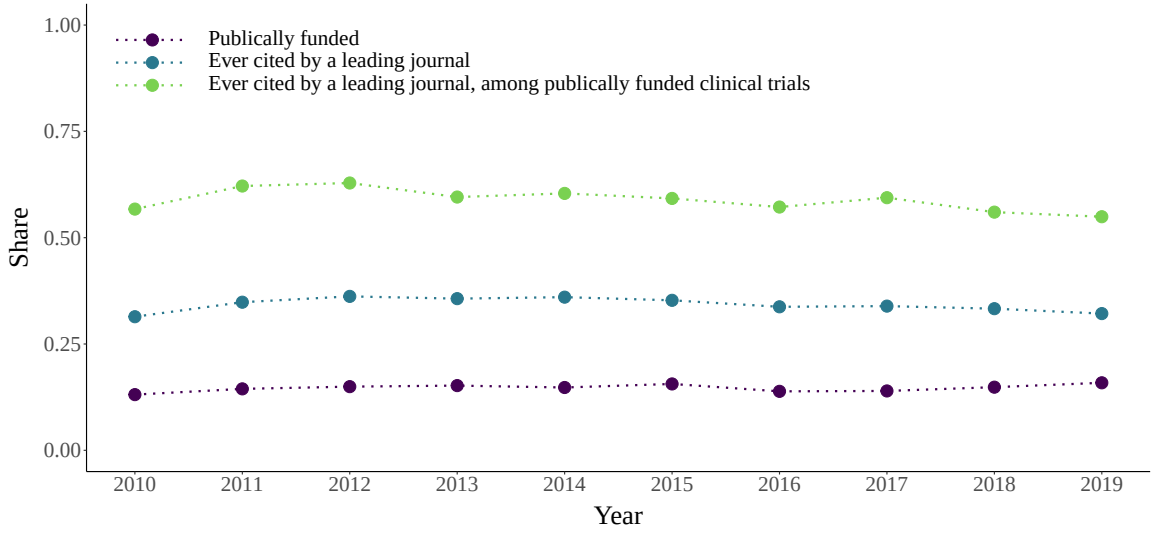
3.2 Clinical Trials as a Productivity Indicator

The trends in Figure 3.1, on their own, allow us to draw two conclusions. First, existing estimates that suggest sharply rising quantities of clinical trials (e.g., Bloom et al., 2020) are a consequence of technical challenges in the measurement of research. Second, existing estimates of rising quantities of medical and scientific research (e.g., Bloom et al., 2020; Chu and Evans, 2021; Park et al., 2023) are *not* driven by shifts in the quantity of clinical trials.

Figure 3.2 documents that—on two additional dimensions—clinical trial production has been stable over the past decade. In particular, Figure 3.2 displays the share of publications in our census with the following characteristics: any funding from a government agency (purple), any (three-year) citations from a leading journal in medicine (teal), or any citations from a leading journal *conditional* on having public funding (green).

We designate a publication as having public funding if it appears in the National Institutes of Health RePORTER data, as linked to a funded research grant, or if PubMed reports a source of U.S.-based research

FIGURE 3.2. Public Funding and Never-Cited Research



Notes: Figure 3.2 displays three time series: the proportion of clinical trials that are publicly funded, the proportion of clinical trials that are ever cited by a leading journal, and the proportion of publicly funded clinical trials that are ever cited by a leading journal. The sample of published clinical trials is constructed with the conservative model.

funding.[30] To designate a set of the "top" journals in medicine, we follow Angrist et al. (2020). We collect all citations originating from the "trunk journals" in medicine, the Journal of the American Medical Association and the New England Journal of Medicine, to records in PubMed between 2010 and 2022. We designate a journal as being "leading" if it received at least 100 citations from a trunk journal over this time period. This yields a list of 84 journals.[31]

Two facts are worth highlighting. First, all three measures are stable, essentially unchanging, over our time period. Second, nearly 70 percent of clinical trials are *never* cited by a leading journal, suggesting substantial, but time-invariant, heterogeneity across trials.

Stability in public funding suggests, on its own, that the allocation of public funding to clinical trials has not considerably shifted over this time period. One can imagine alternative explanations for this stability. For example, perhaps the share of all projects receiving public funding has remained stable, but the cost of each

[30]Technically, PubMed indicates whether a publication received research funding from government agencies and private philanthropies outside of the United States. We restrict consideration to those sources of funding based in the United States because there appears to be a shift in the reporting of certain sources of non-U.S. funding beginning in 2017. The total number of publications linked to a source of public funding increases sharply—by roughly 50 percent from 2016—in 2017. On inspection, other trends are smooth during this time, including trends in public funding for the United States. We are unable to determine if certain sets of authors began to disclose sources of funding in 2017, if certain journals changed their reporting requirements, or if PubMed changed its data construction process. We thus focus on the United States, where we find no evidence of large changes in data reporting practices.

[31]See Appendix A.3 for details. Appendix D reproduces this figure using five-year citations. The trend is unchanged.

trial has increased substantially. Alternatively, perhaps public agencies are funding a higher or lower share of studies, but that these changes are perfectly offset by changes in either study success or publication bias.

We view the first potential explanation as unlikely for two reasons. Somewhat suggestively, NIH RePORTER data allow us to tab total expenditures associated with grants acknowledged in these publications. These grant-based expenditure measures, too, are stable over this period. Note, however, that linking research grant dollars to specific scientific papers is an imprecise exercise. See Li (2017) for an extended discussion of this measurement challenge. More directly, recent work by Sertkaya et al. (2024) suggests that the costs of drug development—measured using the contracts associated with particular study sites in clinical trials—have been essentially constant from 2000 to 2018. The second potential explanation is unlikely from a more conceptual perspective. We find essentially complete stability in this measure over time: any changes in publication bias or trial failure rates must, then, perfectly offset changes in public funding patterns, to yield the flat series here.

Figure 3.3 provides a more granular view of this heterogeneity. We plot quantiles of the distribution of the number of (three-year) citations to publications in our census from publications in leading journals, for each year between 2010 and 2019.[32] Grey areas at the bottom of each plot represent publications that receive zero citations. This distribution is stable over our time period. Observe that we elongate the y-axis to highlight the right-tail of the citation distribution. The quantiles corresponding to the right tail of the citation distribution are constant across years. The left-tail is similarly stable: the share of records receiving zero citations is essentially constant, year-on-year.

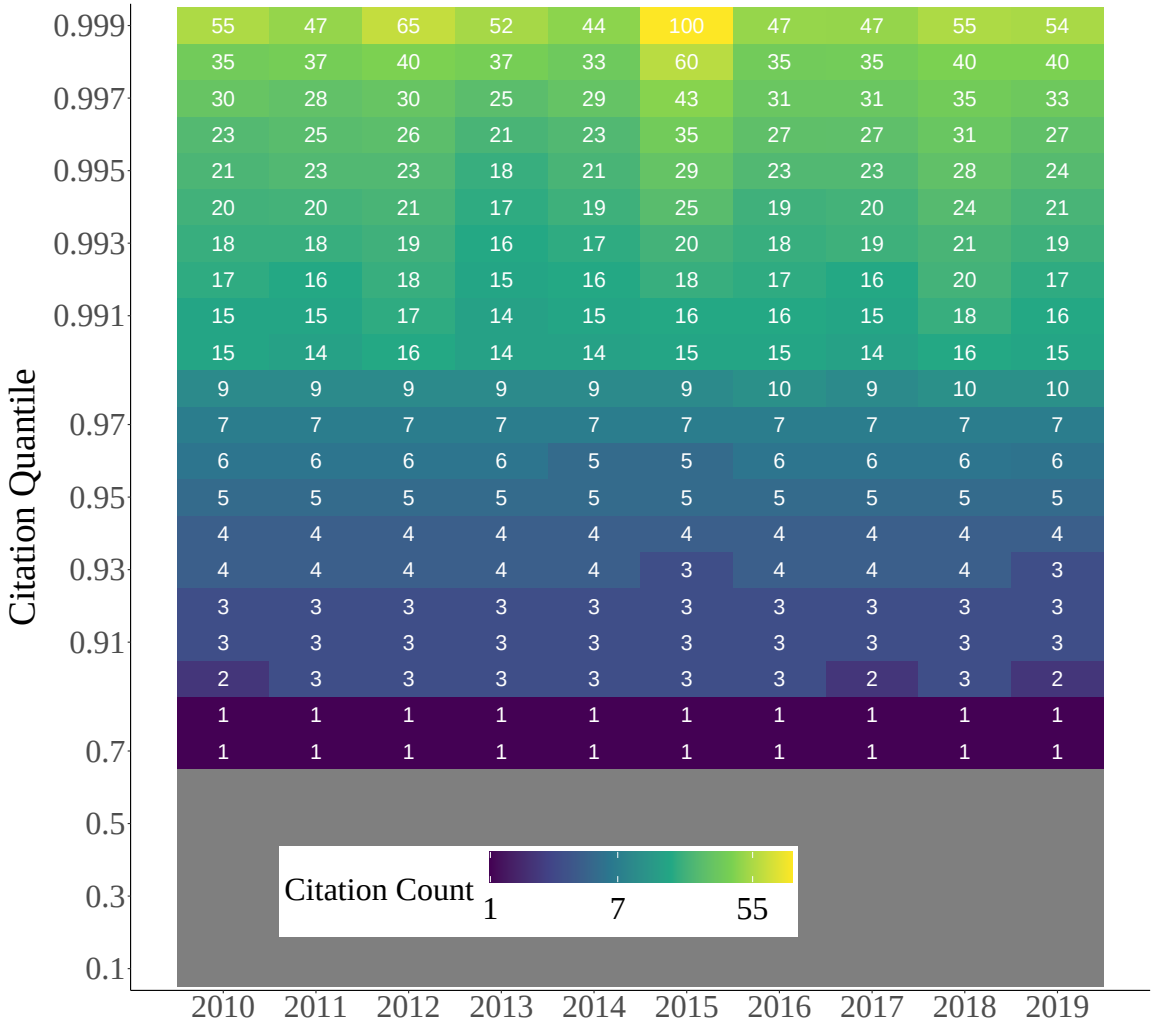
Figure D.4 is constructed analogously, where instead we display quantiles of the distribution of the number of citations received from all publications indexed in PubMed. There, we find that the right-tail of this distribution is increasing over time. In other words, as the total number of publications is increasing, the total number of citations received by clinical trials is increasing. However, these citations are not originating in leading journals.

We interpret this set of facts as evidence that trends in clinical trial production have been stable over this time period. There are four potential concerns with this interpretation.

First, if the costs of either upstream, basic inputs into clinical trials or the costs of clinical trials themselves have increased substantially over this time, the total resources devoted toward research will have increased, even though Figure 3.1 suggests that counts of trials produced are stable. In our view, this is unlikely to be the case. On the issue of clinical trial costs, the most up-to-date, comprehensive evidence on drug development

[32] Appendix D presents these results using alternative cuts of our data and five-year citations. The patterns remain unchanged.

FIGURE 3.3. Citation Distribution Across Time, Clinical Trials



Notes: Figure 3.3 displays a heat map measuring the quantiles of the distribution of citations received by clinical trials in each calendar year from leading journals. The y-axis has been stretched to elongate the right-tail of the citation distribution. Colors are displayed in a log scale. The sample of published clinical trials is constructed with the conservative model.

costs, in our read, is given in Sertkaya et al. (2024), who use internal data from the U.S. Food and Drug Administration and contracts associated with clinical trial sites to compute estimates for the period from 2000 to 2018. In these data, they find that over this period, "the cost of drug development remained relatively stable or may have even decreased."[33] Thus, we do not consider the prospect of rising study prices a concern for the interpretation of this pattern. On the issue of basic science inputs into clinical trials, there is little

[33]This finding revises a frequently cited claim made in popular press interviews (e.g., Herper, 2022), in academic articles that rely on proprietary data sources (e.g., DiMasi et al., 2016), and in surveys on research productivity (Goldin et al., 2024)—which suggest that clinical trial and drug development costs are sharply increasing over time. This revision is in the spirit of an argument made by Cockburn (2006), in response to an earlier debate about drug development costs.

empirical evidence that increases in cost or decreases in the productivity have occurred. A small number of papers examine frictions in markets for basic research. See, for example, Hill and Stein (2021) and Myers (2020). To our knowledge, this literature has not provided evidence of substantial shifts in inputs or factor prices over time, which in turn would increase the cost or decrease the productivity of basic science. In fact, changes in the production of basic scientific inputs around the year 2000—enabled by breakthroughs such as the Human Genome Project and the advent of high-throughput screening for drug discovery (see Scannell et al., 2012)—suggest, if anything, that costs of basic science inputs have fallen. Moreover, it is reasonable that such trends would affect trends in trial quantity, quality, composition, or cost. Given the absence of such patterns, it is unlikely that changes in basic science threaten the interpretation of our findings.[34]

Second, one might be concerned about how to interpret the x-axis in our plots. In particular, the relationship between clinical trial completion dates and clinical trial start dates varies systematically within and across disease classes (Budish et al., 2015). Here, we follow existing literatures that use dates affixed to patents and publications as measures of the timing of research. Practically speaking, as Footnote 14 suggests, records of the start dates of clinical trials are back-filled in administrative databases and, thus, are available only with a considerable lag. We focus on the date of publication, then, for ease of interpretation.

Third, our findings are threatened by publication bias. The existence of publication bias means that counts of scientific publications only partially measure the quantity of research (Andrews and Kasy, 2019). Our main finding is that the quantity of clinical trials is stable over time. If a trend were in fact present, and our measure were confounded by publication bias, then it would need to be the case that the trend was perfectly offset by the bias. It is reasonable to view cancellation, in this way, as unlikely.

Fourth, and finally, we must assume that the number of publications associated with each clinical trial is constant across our period of interest. We view this assumption as reasonable for three reasons. First, by 2010, nearly all scientific journals in our sample had adopted rules requiring authors of clinical trial publications to adhere to a standardized format when reporting trial results (Schulz et al., 2010). Under these rules, there is little scope for authors to, for example, split results across several publications. Second, when constructing our sample of clinical trials, we have excluded publications that re-analyze data generated by an existing study.

[34]Constructing a comprehensive measure of medical research productivity, which grapples with these various input markets, remains an open question. To our knowledge, the work that is closest in spirit is Berndt and Cockburn (2013), who aim to develop a price index for clinical trial research, with the eventual goal of developing price indices for biomedical research more generally.

Third, more suggestively, we find no evidence of a change in this relationship between the size of NIH grants and the quantity of resulting publications reporting the results of a clinical trial during our period of interest.[35]

Taken together, we view the trends in our data, when set against existing empirical evidence, as indicating that clinical trial publication has been stable over this time period.

3.3 Why is the Quantity of Research Rising?

The quantity of papers citing clinical trials has increased, per Figure 3.1, by nearly a factor of two over our period of interest. In this section, we examine various measures of publication composition and quality to shed light on the source of this increase. Our aim is to determine the source of the categorization errors that yield measures such as those in Figure 2.1. In doing so, we provide suggestive evidence on whether measured increases in the quantity of scientific research correspond to "real" increases in the quantity of underlying research or whether they capture changes in the ease, frequency, and composition of publication.

3.3.1 Publication Content. We inspect the technical content of "non-clinical-trial" papers in three ways. We hand-label random samples of 100 records, drawn from the beginning (publication year 2013) and end (publication year 2019) of our sample of papers that cite clinical trials. In this pooled set of 200 records, we identify one record that should have been in our sample of clinical trials.[36] The remainder of the records can be split into seven sets: editorial records (e.g., comments on other papers, errata, editorials, 2.5%), publications reporting preclinical studies (e.g., in-laboratory and in-animal tests, 21.5%), clinical trial protocols (1%), randomized trials studying something other than a drug (2.5%), case studies and case reports (4.5%), observational studies (30.5%), and review articles (e.g., meta-analyses and literature reviews, 37%).[37] We find a 30 percent increase in the total number of review articles between 2013 and 2019, while the quantity of observational studies remains essentially constant (31 in 2013, and 30 in 2019). The increase in the quantity of meta-analyses is consistent with increases documented elsewhere (Ioannidis et al., 2013). Of course, in our small sample of records, it is difficult to draw authoritative conclusions about the role of growth in this type of literature.

We collect two measures of changes in the quantity of meta-analyses, based on the hypothesis that this may be a quantitatively important driver of the changes in our sample. First, we implement a simple search of the

[35]We inspect this relationship by collecting records from the NIH RePORTER database, for years 2010 to 2019. Specifically, we collect records of the total award associated with funded projects and the quantity of any publications flagged by our procedure as clinical trials. For the median NIH grant, we find that an award of roughly $100,000 per clinical trial, and a mean of roughly $200,000 per trial. This is stable across years of data.

[36]The erroneously labeled publication is PMID 23797691, which reports the results of a clinical trial studying a drug to treat HIV. Note that this exercise suggests a false negative rate of 0.005, consistent with estimates reported previously.

[37]In three cases, review articles appeared with the heading "Comment" or "Editorial." We categorize these records as falling into the "Review" category.

abstracts of publications in this "citing" sample. To select keywords for this search, we inspect the set of hand-labeled records, used as the "training data" for our fine-tuning procedure in Section 2.4, as many were flagged as out-of-sample because they were literature reviews or meta-analyses.[38] As expected, given the stringency of our search terms, this yields a small number of records: in our three-year citing trials sample, there are 1,606 records that satisfy these criteria in 2013 and 3,372 in 2019, corresponding to a 109 percent increase.

Our second method generates—as the discussion in Section 2 suggests—what might be viewed as an upper bound on the quantity of meta-analyses and literature reviews in our data. We collect all records indexed in PubMed with the NLM tags "meta-analysis," "systematic review" and "review." There are 9,407 records in our citing sample with one of these tags in 2013 and 14,991 in 2019, corresponding to a 59 percent increase.[39]

From these three procedures, we draw a suggestive conclusion: there has been a quantitatively important increase in the number of meta-analyses over our period of interest. This increase is on the order of 30 to 100 percent, across measurement strategies. The specific quantity of meta-analyses is not of central interest for this paper, except insofar as it allows us to conclude that the quantity of medical research that summarizes existing papers—rather than producing new evidence—appears to be growing over time. To be clear, meta-analyses and similar forms of literature review provide important information to physicians and can be vital in the diffusion of existing information, especially as the stock of knowledge grows (Jones, 2009). This change in composition, however, is relevant context when interpreting measured increases in the quantity of publications.

3.3.2 Publication Geography. Changes in the production of meta-analyses have been linked (e.g., Ioannidis et al., 2013) to shifts in the geographic distribution of medical research. Publication data allow us to inspect geographic trends in our sample.

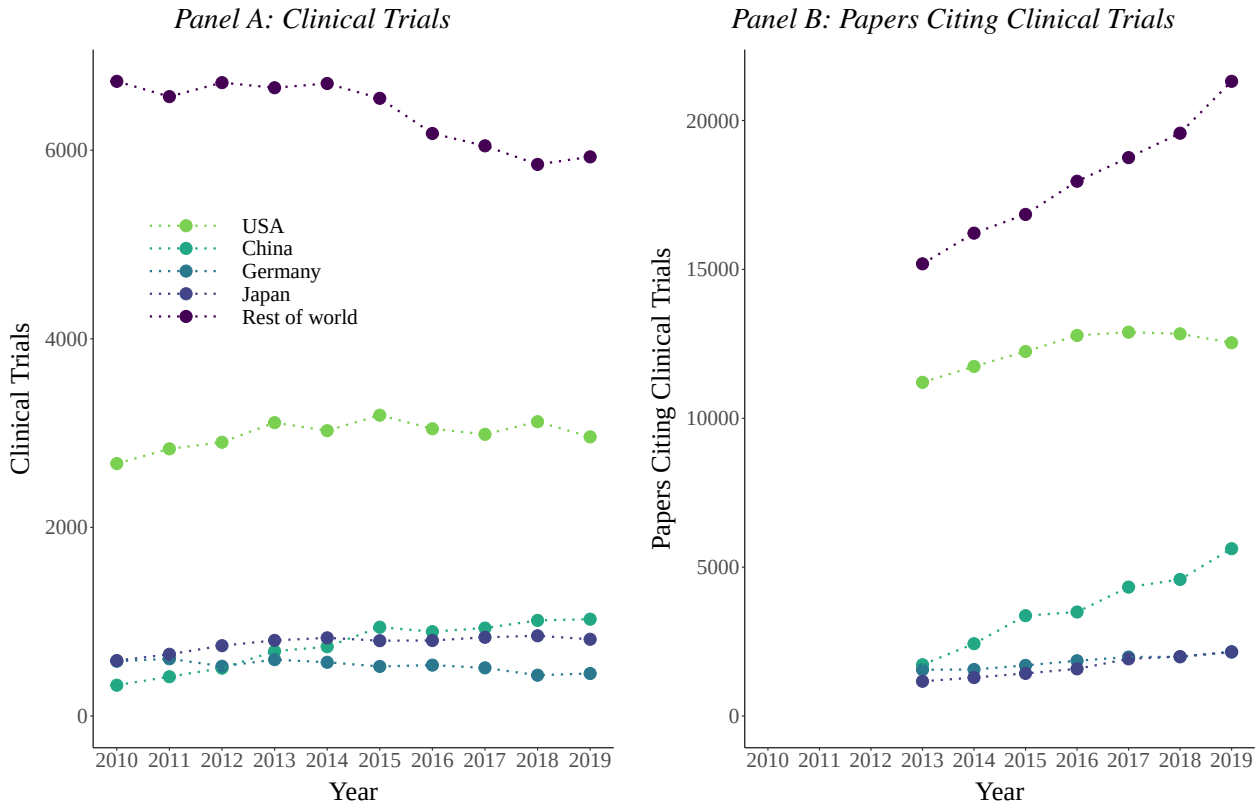
Figure 3.4 displays trends in the annual quantity of published clinical trials, and papers citing clinical trials, disaggregated by the location of the first-listed author.[40] In our sample, the top four producers of published clinical trials, and papers citing clinical trials, are, in order, the United States, China, Germany, and Japan.

[38]We searched for the following keywords: meta-analysis, metaanalysis, metaanalyses, systematic review, systematic reviews, systematically review, systematic search, review of published data, literature review, literature search, search of databases, review all literature, reviewed all literature, narrative review, systemic review. We also included any record that mentioned at least two of the following databases, which—based on our hand-labeling experience—are often referenced as part of the methods involved in a literature review or meta-analysis: MEDLINE, EMBASE, CINAHL, PubMed, Cochrane Central Register of Controlled Trials, BioMedCentral.

[39]Figure D.8 displays a time series of the number of 'citing' papers that satisfy both our keyword and NLM-tag based methods for identifying meta-analyses.

[40]In most domains of scientific and medical publication, last-listed authors are senior investigators. First-listed authors are typically junior investigators. Agha and Molitor (2018) observe that, for large-scale clinical trials, first-authors are more likely to be the principal investigator. In our context, the first- and last-author have the same listed country for 85 percent of clinical trial records and 91 percent of non-clinical trial records. We collect details on author location from Clarivate Analytics' Web of Science. See Appendix A.4 for further details.

FIGURE 3.4. Composition of Publications Across Countries



Notes: Figure 3.4 displays measurements of the number of clinical trials, and papers that cite clinical trials, published in each calendar year, by country. The sample of published clinical trials is constructed with the conservative model. To address truncation, we report the number of publications that cite clinical trials published in the preceding three years.

Clinical trial quantities—across geographies—are stable over our period of interest.[41] Roughly 30 percent of published clinical trials originate in the United States. This share is unchanging across periods. By contrast, Panel B indicates a large shift in the medical research ecosystem. Since 2013, there has been a small increase in the quantity of other medical research published by authors in the United States. This small change is dwarfed by increases—on the order of 30-230 percent—in the quantity of other medical research from China and from countries outside of the top four producers of medical research.

Institutional details are helpful in making sense of the differential patterns across these cuts of data. Regulatory requirements and local capabilities are key factors influencing where clinical trials are conducted and published. Sites must be capable of producing data sufficient to persuade regulators in high-value

[41]Note that Panel A of Figure 3.4 displays measurements of the location of clinical trial publication authors, not the location of clinical trial sites. We cannot rule out that, for example, trials authored by researchers with United States mailing addresses were conducted elsewhere. Thus, we do not interpret these as facts on clinical trial "offshoring." See Petryna (2007) and Durvasula (2023) for longer discussions of the geography of clinical trial *sites*.

markets. Certain regulators require that sponsors of new drug applications submit trial evidence collected from a domestic population.[42] In recent years, the U.S. Food and Drug Administration (FDA)—responding to concerns from patient groups—has indicated that it may not be willing to approve drugs on the basis of evidence collected from exclusively foreign sites (for a longer discussion, see Alsan et al., 2024). In one especially high-profile example, a cancer drug was rejected by the FDA after being tested exclusively in China (Kolata, 2022). As long as the U.S. market continues to have outsized value for pharmaceutical firms—a consequence of the especially high prices paid by American consumers—it may be unsurprising to find that trial evidence disproportionately originates in the United States, even as global production of other forms of medical research continues to grow. Observational studies, meta-analyses, and case reports are, by contrast, cheap and not subject to the same financial and regulatory pressures. Digital tools, including search engines and the large language models deployed in this paper, will likely further depress these costs, potentially further widening this gap.

Four countries are the most prolific producers of clinical trials and other forms of medical research. Of interest here, however, are those countries that experienced the largest growth in their production of medical research over this period. On this measure, China, Poland, Japan, and Spain top the ranking in our data. In Appendix D.2, we provide more details on the set of countries with the largest growth in this form of medical research over our period of interest. Several details are worth noting. First, Chinese production of records in our citing sample increases by roughly 225 percent, producing 5,620 papers in our citing set in 2019. The next country in this ranking, Poland, experiences 96 percent growth and produces 311 papers in our citing set in 2019. The top 23 countries in terms of growth, listed in the appendix, produce 93 percent of all 2019 papers in this set. Per the World Bank income classification scheme, nearly every country in this fast-growing group is high income. The only countries categorized differently are China, India, Brazil, Iran, and Turkey, which the World Bank classifies as either upper (China, Brazil, Iran, Turkey) or lower (India) middle-income.[43] That is, this trend captures, in part, an increase in the quantity of diffusion-focused literature in high-income countries that have long had robust health infrastructure. Below, we establish that on several quality measures, these publications appear to be declining in value over time—consistent with the idea that some of this increase in production may be a response to changing scientific incentives for publication.

[42]Until December 2023, Japan required domestic phase I clinical trials of drugs developed overseas before Japanese individuals could participate in international phase III trials for pharmaceutical regulatory approval. This policy reflected, in part, differences in disease burden in Japan, relative to countries where drugs are routinely tested. This may explain why Japan is a top producer of clinical trials over our sample period. See Namba et al. (2024) for one discussion.

[43]https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html

3.3.3 Publication Quality. Next, we collect a set of facts about the quality of this growing set of "citing" papers. We begin in Figure 3.5 by replicating Figure 3.3 for the sample of papers that cite clinical trials. There are several elements of this plot that are informative. First, the share of "citing" papers with a small number of citations appears to grow over time. That is—the left tail of the citation distribution for this set is both long and growing. Note also that the right tail of the citation distribution appears to shrink over time. That is, the proportion of highly cited papers appears to be shrinking. These patterns provide one indication that the central tendency of the quality of non-clinical-trial medical research may be decreasing over time.[44]

Citations are imperfect measures of the ex post quality of a scientific publication. Our data allow us to examine several other measures that provide suggestive insight into the ex ante quality of these publications. Unfortunately, shifting geographic trends in this sample and the low quality of measures of public funding for countries outside of the United States render it difficult to use a measure of public funding as a proxy for quality. We do find a 10 percentage point decrease in the share of publications in this set that record having received (U.S.-based) public funding, but this is likely best interpreted as a mechanical consequence of the geographic heterogeneity displayed in Figure 3.4.

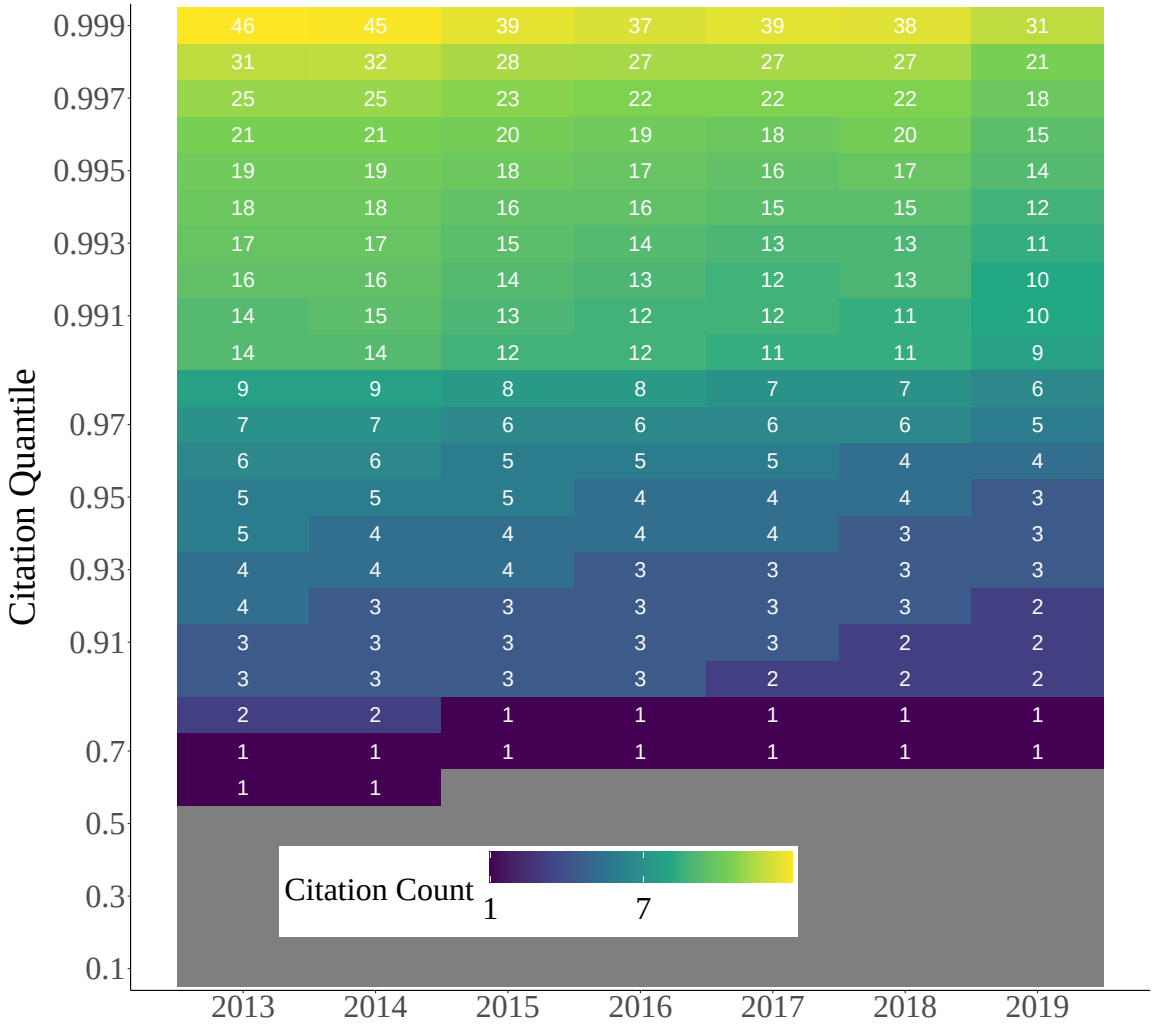
Instead, we focus on author characteristics, constructed using publication data, that also shed some light on research quality. Scientific author norms typically mean that first-listed authors are junior scientists and last-listed authors are senior, lead scientists. We collect facts about these groups separately.

We begin by examining the productivity of lead scientists who authored papers in our citing sample. We collect all publications indexed in PubMed authored by lead scientists (last-authors) who write at least one paper in our citing sample. The median scientist in this set writes five publications per year. We observe opposing trends in annual publication quantity and citation patterns during this time. We find that the median quantity of papers written by a scientist per year increases from 3 to 5 over this period. The median number of citations to those papers from leading journals drops, over the same period, from 3 to 1. At the 90th percentile, the quantity of per scientist publications per year is stable, at 9 per scientist between 2013 and 2019. Citations from leading journals are similarly stable, at 10 per publication.[45] For the median senior scientist who appears in our citing sample in 2013, approximately 80 percent of their previously published papers (i.e., published before 2013) had received at least one citation. For scientists who first appear in our citing sample in 2019, the share of previous publications with a non-zero number of citations falls to roughly 70 percent.

[44] Figure D.5 gives an analogous figure, displaying the distribution in citations from all publications to the sample of publications that cite clinical trials. Here, the right tail of the distribution of citations is growing.

[45] We observe sharp declines in the likelihood that a scientist has ever received funding from the NIH or from another U.S. federal agency. As noted above, however, we hesitate to place much weight on this fact, as it likely reflects changes in the composition of the sample.

FIGURE 3.5. Citation Distribution Across Time, Papers that Cite Clinical Trials



Notes: Figure 3.5 displays a heat map measuring the quantiles of the distribution of citations received by papers that cite clinical trials, in each calendar year, from leading journals. The y-axis has been stretched to elongate the right-tail of the citation distribution. Colors are displayed in a log scale. The sample of published clinical trials is constructed with the conservative model. The sample of papers that cite clinical trials is constructed using a three-year window.

For junior scientists, we focus on a measure of attrition: after authoring a paper in our citing sample, do we ever observe another publication written by that junior scientist?[46] We collect all publication records in the Web of Science associated with these junior (first-listed) authors. In 2013, the likelihood that a junior scientist published another paper in 2014 was roughly 15 percent. By 2019, this falls to roughly 8 percent. We observe similar trends when two- and three-year publication patterns are observed. We interpret this

[46] Hill and Stein (2021) use a similar measure of attrition, when considering the impacts of losing a priority race for scientists' careers.

attrition as evidence that junior scientists who author papers in our citing sample may be less likely to remain research productive in the future.

Taken together, we interpret these facts about scientist research productivity as suggestive evidence of declining average research quality in our citing sample over time. When interpreted in light of the citation distribution trends in Figure 3.5, we conclude that measured increases in the quantity of this "other" medical research may be accompanied by a decline in its average quality.

4. Discussion

Studies of innovation and productivity have long grappled with the prospect of diminishing returns to research (e.g., Merton, 1935; Stafford, 1952; Schmookler, 1952, 1954, 1966; Bloom et al., 2020). Answers to the question at the center of this literature—is the cost of sustained technological progress increasing over time?—have been unsatisfying, as fundamental challenges in the measurement of technological innovation have limited our ability to do more than determine that available metrics do not yield conclusive answers (Griliches, 1990, 1994). In one sense, this paper fits in this tradition, in that we find that, in the context of pharmaceutical research, recent concerns about declining productivity are an artifact of confounded measurements of research quantities.

On the other hand, enabled by advances in generative AI, we develop a procedure for labeling unstructured text that allows us to produce data with an accuracy, precision, and scale sufficient to make progress on the measurement challenges inherent to empirical studies of innovation. We apply this procedure to identify records of clinical trials from a corpus of scientific publications. The resulting census corrects errors in previous measures of clinical trial research. We show that trends in the quantity, quality, and composition of clinical trials have been stable since 2010.

We contrast these trends with analogous patterns in the production of other forms of medical research. We find that, since 2010, the quantity of other forms of medical research has roughly doubled, a pattern consistent with estimates reported in other papers. On inspection, the increase in quantity is driven by a large increase in the number of papers authored by scientists in China and in the number of papers that synthesize existing literature (e.g., literature reviews and meta-analyses). Growth in these forms of research coincides with a decline in various measures of average publication quality. These facts provide evidence on the source of classification errors in existing measures of clinical trial production and, more generally, provide context on compositional trends relevant to studies that use publications as measures of research.

These patterns are, in many ways, unsurprising. In recent years, editorials in prominent medical journals (Harvey, 2020) have observed that changes in the incentive structures for physicians, scientists, and even

medical students, across countries, have led to "an extreme of quantity at the expense of quality" in medical publishing (Siegel et al., 2018).[47] Recent work documents a substantial increase in the quantity of meta-analyses and systematic reviews—summaries of existing clinical trials—driven by scientists outside of the United States (Ioannidis, 2016). In parallel, there has been a rapid increase in the number, and scale, of "mega-journals"—scientific journals published frequently, with a large and growing number of pages (Ioannidis et al., 2023).

Despite these changing incentives, substantial constraints—financial, ethical, and practical—keep firms, physicians, and scientists from increasing the quantity of clinical trials. Clinical trials are infrastructure-intensive. Trials are run at specific sites, which must identify patients who satisfy all relevant eligibility criteria and monitor them for periods ranging from six months to twenty years (Piantadosi, 2024). Sites may struggle to find patients who meet these criteria, especially in small disease markets (Kolata, 2017), and the costs of accruing large samples for certain studies can be prohibitive (Alsan et al., 2024).

Our findings focus on one aspect of the larger biomedical research ecosystem: investments in efforts to bring basic science discoveries to market, in the form of new medicines. Development of supplemental measures of medical research productivity, which capture the efficiency of basic scientific research, is a useful area for future research.

Our findings, also, apply to roughly one decade. We impose this restriction because a set of policy changes in the early 2000s substantially shifted incentives for disclosure and publication of clinical trial results (Laine et al., 2007; Schulz et al., 2010). To be clear, there is nothing about our method that limits its applicability to data reported in the particular form required by these policy changes. Our concern here is about the interpretability of the resulting series. As noted previously, proxies for scientific research provide very little insight if the relationship between one unit of the proxy and units of underlying, unobserved research shift over time. Policy changes in this sector are such that we risk introducing this source of bias with a longer series. Thus, we cannot speak to trends that precede this period. From the perspective of contemporary public policy, however, trends from 2010 to the present are likely to be most relevant.

Our revision of estimates of clinical trial production is both, from a policy perspective, optimistic and pessimistic. On the one hand, we find evidence consistent with a market in equilibrium: clinical trial production, on a variety of margins, appears stable. This directly calls into question the stylized fact of

[47]See, as one example of how these incentives have shifted, program certification requirements from the Accreditation Council for Graduate Medical Education in Accreditation Council for Graduate Medical Education (2017). The ACGME provides formal accreditation for training programs and, as part of this process, lays out expectations of minimum quantities of publication for medical students and clinical faculty. Note that certain researchers have put forth alternative explanations for the rapid increase in the total number of publications. A trend toward evidence-based decision-making in medicine has increased demand for articles that synthesize large bodies of research (Lohr, 2004).

diminishing returns to drug development, which is viewed in some settings (e.g., Scannell et al. (2012) and Goldin et al. (2024)) as an empirical regularity on par with Moore's Law for semiconductor development.

On the other hand, it is worth remarking, briefly, on the implications of completely stable clinical trial trends. The early twenty-first century is often described by scientists as a "golden age" for medicine. Breakthroughs, such as the completion of the Human Genome Project and the diffusion of DNA sequencing technologies one billion times faster than their predecessors, rapidly expanded the set of available scientific opportunities (e.g., Scannell et al., 2012). Technological advances were matched by institutional support, including the doubling of the budget of the U.S. National Institutes of Health between 1998 and 2003. Against these factors that seemingly forecast a period of rapid innovation, the facts documented in this paper—stability in the publication of clinical-trials, but rising quantities of other forms of translation research—highlight the importance of understanding frictions that affect the translation of basic scientific insights into medical technologies relevant to patients.

Specifically, these patterns suggest that there is likely value in efforts to increase the productivity and decrease the costs of clinical trials, and in efforts to reconcile incentives in academic publishing with social value. A long-standing discussion in medicine and statistics, spurred by Altman (1994), centers on the argument that "we need less research, better research, and research done for the right reasons." Our data indicate that nearly half of all clinical trials published in our sample are never cited by leading medical journals, and roughly 15 percent are never cited at all.

REFERENCES

Accreditation Council for Graduate Medical Education (2017). ACGME common program requirements. Accessed: 2024-05-05.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agha, L., Kim, S., and Li, D. (2022). Insurance design and pharmaceutical innovation. *American Economic Review: Insights*, 4(2):191–208.

Agha, L. and Molitor, D. (2018). The local influence of pioneer investigators on technology adoption: Evidence from new cancer drugs. *Review of Economics and Statistics*, 100(1):29–44.

Alsan, M., Durvasula, M., Gupta, H., Schwartzstein, J., and Williams, H. (2024). Representation and extrapolation: Evidence from clinical trials. *The Quarterly Journal of Economics*, 139(1):575–635.

Altman, D. G. (1994). The scandal of poor medical research. *Bmj*, 308(6924):283–284.

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2020). Inside job or deep impact? extramural citations and the influence of economic scholarship. *Journal of Economic Literature*, 58(1):3–52.

Bartik, A., Gupta, A., and Milo, D. (2023). The costs of housing regulation: Evidence from generative regulatory measurement. *Available at SSRN*.

Berndt, E. R. and Cockburn, I. M. (2013). Price indexes for clinical trial research: a feasibility study. Technical report, National Bureau of Economic Research.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Brainard, J. (2020). Dozens of scientific journals have vanished from the internet, and no one preserved them. *Science, September*, 8:2020.

Broad, W. J. (2023). What happened to all of science's big breakthroughs? *The New York Times*.

Bryan, K. A. and Williams, H. L. (2021). Innovation: Market failures and public policies. *Handbook of industrial organization*, 5(1):281–388.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Budish, E., Roin, B. N., and Williams, H. (2015). Do firms underinvest in long-term research? evidence from cancer clinical trials. *American Economic Review*, 105(7):2044–2085.

Bybee, J. L. (2023). The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*.

Chavez-MacGregor, M. and Giordano, S. H. (2016). Randomized clinical trials and observational studies: Is there a battle? *Journal of Clinical Oncology*, 34(8):772–773.

Chertman, W. (2023). Slow, costly clinical trials drag down biomedical breakthroughs. Issue brief, Manhattan Institute.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Chu, J. S. and Evans, J. A. (2021). Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118.

Cockburn, I. M. (2006). Is the pharmaceutical industry in a productivity crisis? *Innovation policy and the economy*, 7:1–32.

Collison, P. and Nielsen, M. (2018). Science is getting less bang for its buck. *The Atlantic*.

Cortellis (2024). Cortellis Labs. https://www.cortellislabs.com/page/?api=api-CLI. Accessed on: 2024-05-04.

Cunningham, C., Ederer, F., and Ma, S. (2021). Killer acquisitions. *Journal of political economy*, 129(3):649–702.

Dell, M. (2024). Deep learning for economists. *arXiv preprint arXiv:2407.15339*.

DeVito, N. J., Bacon, S., and Goldacre, B. (2020). Compliance with legal requirement to report clinical trial results on clinicaltrials. gov: A cohort study. *The Lancet*, 395(10221):361–369.

DeVito, N. J. and Goldacre, B. (2021). Evaluation of compliance with legal requirements under the fda amendments act of 2007 for timely registration of clinical trials, data verification, delayed reporting, and trial document submission. *JAMA Internal Medicine*, 181(8):1128–1130.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of health economics*, 47:20–33.

Durvasula, M. M. (2023). Inclusive infrastructure for clinical trials. *Brookings: Building a Better NIH*.

Durvasula, M. M., Ouellette, L. L., and Williams, H. L. (2021). Private and public investments in biomedical research. In *AEA Papers and Proceedings*, volume 111, pages 341–345. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Executive Office of the President (2019). Launch of the national science and technology council joint committee on improving research environments. Memo. The White House Unveils New Joint Initiative Supporting U.S. Research Community.

Feldman, S., Ammar, W., Lo, K., Trepman, E., van Zuylen, M., and Etzioni, O. (2019). Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA network open*, 2(7):e196700–e196700.

Finkelstein, A. (2004). Static and dynamic effects of health policy: Evidence from the vaccine industry. *The Quarterly Journal of Economics*, 119(2):527–564.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Goldin, I., Koutroumpis, P., Lafond, F., and Winkler, J. (2024). Why is productivity slowing down? *Journal of Economic Literature*, 62(1):196–268.

Graboyes, R. and Reimschisel, J. (2017). Insurance costs just the tip of the iceberg. Time to reform the FDA. *The Hill*. Opinion contributors.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, 28(4).

Griliches, Z. (1994). Productivity, r&d, and the data constraint. *American Economic Review*, 84(1):1–23.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Guzman, J., Jelveh, Z., and Kogut, B. (2024). The unfolding spirit of science: Evidence from p-hacking in academic clinical trials. Technical report, University of Maryland mimeo.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2000). Market value and patent citations: A first look.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights and methodological tools.

Harvey, L. A. (2020). We need to value research quality more than quantity. *Spinal cord*, 58(10):1047–1047.

Herper, M. (2022). Here's why we're not prepared for the next wave of biotech innovation. *STAT*. Accessed: 2022-11-03.

Hill, R. and Stein, C. (2021). Race to the bottom: Competition and quality in science. *Northwestern University and UC Berkeley*.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Inflation Reduction Act (2022). Public Law No. 117-169, 136 Stat. 1818. United States Congress.

Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3):485–514.

Ioannidis, J. P., Chang, C. Q., Lam, T. K., Schully, S. D., and Khoury, M. J. (2013). The geometric increase in meta-analyses from china in the genomic era. *PloS one*, 8(6):e65602.

Ioannidis, J. P., Pezzullo, A. M., and Boccia, S. (2023). The rapid growth of mega-journals: Threats and opportunities. *Jama*, 329(15):1253–1254.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Jones, D. S. and Podolsky, S. H. (2015). The history and fate of the gold standard. *The Lancet*, 385(9977):1502–1503.

Kao, J. (2024). Information disclosure and competitive dynamics: Evidence from the pharmaceutical industry. *UCLA mimeo*.

Kao, J. and Oostrom, T. (2024). Representation in product development: Evidence from insurance and clinical trials. *UCLA mimeo*.

Kao, J., Ross, J. S., and Miller, J. E. (2023). Transparency of results reporting in cancer clinical trials. *JAMA Network Open*, 6(8):e2328117–e2328117.

Kolata, G. (2017). A cancer conundrum: too many drug trials, too few patients. *The New York Times*, 12.

Kolata, G. (2022). FDA Panel Rejects Lilly's Cancer Drug Tested Only in China. *The New York Times*.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Laine, C., Horton, R., DeAngelis, C. D., Godlee, F., Drazen, J. M., Frizelle, F. A., Haug, C., Hébert, P. C., et al. (2007). Update on trials registration: Clinical trial registration: Looking back and moving ahead. *International Committee of Medical Journal Editors*.

Lerner, J. and Seru, A. (2022). The use and misuse of patent data: Issues for finance and beyond. *The Review of Financial Studies*, 35(6):2667–2704.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics*, 9(2):60–92.

Lichtenberg, F. R. (2018). The impact of biomedical research on US cancer mortality. *Measuring and Modeling Health Care Costs*, 76:475.

Liu, T. and Low, B. K. H. (2023). Goat: Fine-tuned LLaMA outperforms GPT-4 on arithmetic tasks. *arXiv preprint arXiv: 2305.14201*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lohr, K. N. (2004). Rating the strength of scientific evidence: relevance for quality improvement programs. *International Journal for Quality in Health Care*, 16(1):9–18.

Long, T. (2023). The hidden toll of drug price controls: Fewer new treatments and higher medical costs for the world. Policy brief, Information Technology & Innovation Foundation.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Merton, R. K. (1935). Fluctuations in the rate of industrial invention. *The Quarterly Journal of Economics*, 49(3):454–474.

Moss, E., Nunn, R., and Shambaugh, J. (2020). The slowdown in productivity growth and policies that can restore it. *The Hamilton Project, Brookings Institution, Washington, DC*.

Mullainathan, S. and Rambachan, A. (2024). From predictive algorithms to automatic generation of anomalies. Technical report, National Bureau of Economic Research.

Myers, K. (2020). The elasticity of science. *American Economic Journal: Applied Economics*, 12(4):103–134.

Namba, M., Kaneda, Y., Ozaki, A., and Tanimoto, T. (2024). Clinical trials: Japan's opt-out policy raises risks of adverse drug responses. *Nature*, 626(7998):261–261.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Oostrom, T. (2024). Funding of clinical trials and reported drug efficacy. *Journal of Political Economy*, Forthcoming.

Ouellette, L. L. and Sampat, B. N. (2024). The feasibility of using Bayh-Dole march-in rights to lower drug prices: An update. Technical report, National Bureau of Economic Research.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pakes, A. (1986). Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica: Journal of the Econometric Society*, pages 755–784.

Pakes, A. and Griliches, Z. (1980). Patents and r&d at the firm level: A first report. *Economics letters*, 5(4):377–381.

Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. (2023). Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*.

Petryna, A. (2007). Clinical trials offshored: On private sector science and public health. *BioSocieties*, 2(1):21–40.

Piantadosi, S. (2024). *Clinical trials: a methodologic perspective*. John Wiley & Sons.

Piper, K. (2023). Why is science slowing down? *Vox*.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Roy, A. (2021). High drug prices don't accelerate innovation—lower r&d costs do. *FREOPP.org*.

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

Scherer, F. (1984). Using linked patent and r&d data to measure interindustry technology flows. In *R&D, patents, and productivity*, pages 417–464. University of Chicago Press.

Schmookler, J. (1952). The changing efficiency of the american economy, 1869-1938. *The review of Economics and Statistics*, pages 214–231.

Schmookler, J. (1954). The level of inventive activity. *The Review of Economics and Statistics*, pages 183–190.

Schmookler, J. (1966). *Invention and economic growth*. Harvard University Press.

Schulz, K. F., Altman, D. G., and Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and pharmacotherapeutics*, 1(2):100–107.

Sertkaya, A., Beleche, T., Jessup, A., and Sommers, B. D. (2024). Costs of drug development and research and development intensity in the us, 2000-2018. *JAMA Network Open*, 7(6):e2415445–e2415445.

Siegel, M. G., Brand, J. C., Rossi, M. J., and Lubowitz, J. H. (2018). "Publish or perish" promotes medical literature quantity over quality. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 34(11):2941–2942.

Stafford, A. B. (1952). Is the rate of invention declining? *American journal of sociology*, 57(6):539–545.

Stephenson, J. (2021). In a first, fda warns company to remedy failure to post clinical trial results. *JAMA Health Forum*, 2(5):e211306–e211306.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

The Economist (2022). How to escape scientific stagnation. *The Economist*. Finance & Economics — Free Exchange.

Thomas, J., McDonald, S., Noel-Storr, A., Shemilt, I., Elliott, J., Mavergames, C., and Marshall, I. J. (2021). Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for cochrane reviews. *Journal of Clinical Epidemiology*, 133:140–151.

Thompson, D. (2021). America is running on fumes. *The Atlantic*.

Trinh, T. H. and Le, Q. V. (2019). A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Tse, T., Fain, K. M., and Zarin, D. A. (2018). How to avoid common problems when using clinicaltrials. gov in research: 10 issues to consider. *Bmj*, 361.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wikimedia Foundation (2023). Wikimedia downloads. *https://dumps.wikimedia.org*.

Xu, C., Guo, D., Duan, N., and McAuley, J. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., and Zhou, T. (2024). A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

York, E. (2023). Inflation reduction act's price controls are deterring new drug development. Policy brief, Tax Foundation. 5 min read.

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2021). Big Bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*.

Zarin, D. A., Tse, T., Williams, R. J., and Rajakannan, T. (2017). Update on trial registration 11 years after the icmje policy was established. *New England Journal of Medicine*, 376(4):383–391.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*.

Supplemental Appendix to:

Distilling Data from Large Language Models:

An Application to Research Productivity Measurement*

Maya M. Durvasula⋆ Sabri Eyuboglu† David M. Ritzwoller‡

Contents

Appendix A. Publication Data 1
A.1. The PubMed / MEDLINE Database 1
A.2. The Universe of Potential Clinical Trials 2
A.3. Determining the Set of Leading Medical Journals 5
A.4. The Web of Science 6
Appendix B. ClinicalTrials.gov 7
Appendix C. Prompt Design, Fine-Tuning, and Performance Assessment 10
C.1. Hand-Labeling 11
C.2. Benchmark Comparison to Standard Machine Learning Methods 11
C.3. Prompt Design and Error Analysis 12
C.4. Error Analysis 16
Appendix D. Additional Figures and Further Analyses 16
D.1. Robustness of Trends 16
D.2. Growth in Citing Papers by Geography 24
D.3. Quantifying Meta-Analyses 24
Appendix E. Prompt Repository 28

Date: September 11, 2024
⋆ Stanford Department of Economics and Stanford Law School, maya.durvasula@stanford.edu
† Stanford Department of Computer Science, eyuboglu@stanford.edu
‡ Stanford Graduate School of Business, ritzwoll@stanford.edu

APPENDIX A. PUBLICATION DATA

In this Appendix, we detail our treatment of the PubMed / MEDLINE database (hereafter, "PubMed"). We give a general overview of the PubMed data in Appendix A.1. In Appendix A.2, we describe the process of identifying our initial sample of abstracts, i.e., the "Potential Universe" displayed in Figure 3.1. We discuss the process of determining the set of leading medical journals in Appendix A.3.

A.1 The PubMed / MEDLINE Database

Our version of the PubMed / MEDLINE database contains roughly 34 million records and is current through December 2022. These records were constructed by parsing bulk MEDLINE XML files.[48] Technically, PubMed and MEDLINE are different products. MEDLINE, a subset of PubMed, is the U.S. National Library of Medicine's (NLM) bibliographic database, which contains references to journal articles in life sciences, with a primary focus on biomedicine. A committee at the NLM determines the set of indexed journals, meaning that only journals that meet certain quality and content standards are indexed. In practice, this means that so-called predatory journals are excluded, as are pre-prints and non-peer reviewed articles. Informal conversations with staff at the NLM suggest that MEDLINE *should* contain the universe of peer-reviewed publications in legitimate journals. PubMed includes a broader set of records, including pre-prints and publications deposited through alternative processes. Official documentation for MEDLINE is available at the link: `https://www.nlm.nih.gov/databases/download/pubmed_medline_documentation.html`. Throughout the text, and from this point forward in the Appendix, we refer to this database as "PubMed."

A natural question for researchers who rely on these data as a census of research investments is whether PubMed records are complete and accurate. To our knowledge, there is no paper that reports such validation exercises for each field of PubMed. We focus on those data elements most relevant to our work. There are 2,335,653 records in PubMed that have no associated year of publication, which we drop from our baseline sample. There are 16 records, across all years of data, that are not indexed with tags describing their contents (NML tags). From 2010 forward, roughly 92 percent of records have associated abstracts. We randomly inspect roughly 150 records with no abstracts. None correspond to publications that satisfy our definition of a clinical trial.

We assume, throughout this paper, that records with missing citation data have zero associated citations. We confirm that this is generally accurate in two ways. First, we search for Google Scholar records corresponding to roughly 100 randomly selected records with zero citations. In each case, Google Scholar indicates that the paper has no more than two citations. Second, we link records in PubMed to Clarivate Analytics' Web of Science database. We compare citation counts constructed using information in the two databases. We find that the two measures have a correlation of 0.9.

[48] We are grateful to Heidi Williams for sharing this processed data.

Researchers interested in different cuts of the data may need to conduct additional validation exercises. Figure 1 highlights that the frequencies of certain NLM tags have changed over time, often sharply. Researchers should account for such changes in any use of these flags. For other assessments of the completeness of certain PubMed fields, see Durvasula et al. (2021) and Ouellette and Sampat (2024).

A.2 The Universe of Potential Clinical Trials

We construct our sample using records drawn from the universe of publications indexed in PubMed. The initial sample consists of 34,957,127 unique publications indexed in PubMed, as of 15 April 2023. We drop 2,335,653 records that are missing information on publication year, to yield a base sample of 32,621,474 records. From 2010 forward, at least 92 percent of scientific publications published in each year have associated abstracts. We also drop 14,179 publications with publication year 2023, as we have incomplete data for 2023.

A.2.1 NLM Tags. The National Library of Medicine (NLM) assigns each publication a 'pubtype.' In the entirely of PubMed, there are 16 records missing pubtype tags. To the best of our knowledge, there have been no efforts to validate the PubMed indexing process used to generate these flags. We flag records with each of the following pubtype (or associated unique identifiers, reported in the field 'pubtypeUI') tags:

> Adaptive trial; Clinical conference; Clinical study; Clinical trial; Clinical trial protocol; Clinical trial, Phase 1; Clinical trial, Phase 2; Clinical trial, Phase 3; Clinical trial, Phase 4; Comparative study; Controlled clinical trial; Equivalence trial; Evaluation study; Observational study; Pragmatic clinical trial; Randomized controlled trial; Twin study; Validation study.

These categories are chosen to include all pubtypes with the potential to contain a publication satisfying the restriction enumerated in Definition 2.1. In particular, we follow Feldman et al. (2019), who use multiple pubtype fields to retrieve a sample of records. Here, two authors reviewed the list of potential pubtypes to identify categories likely to include records of interest. Table A.1 reports the frequency of each NLM tag across all records in PubMed and subset to those published after 1 January 2010.

A.2.2 Clinical Trial Registry Identifiers. In 2004, the International Committee of Medical Journal Editors recommended that research journals decline to publish outcomes associated with trials not pre-registered in some repository. Laine et al. (2007) summarizes these policies, based on a 2007 revision, in more detail. The U.S. Food and Drug Administration Amendments Act of 2007, Section 801, mandates registration of all clinical trials regulated by the FDA in ClinicalTrials.gov. Many countries have adopted similar guidance. Several countries and international organizations now maintain registries of trials.

Registry identifiers are distinctive strings of letters and numbers. We flag records that include clinical trial registry identifiers in their abstract text. In particular, we search for records containing acronyms associated with the following registries:

TABLE A.1. Composition of PubMed by NLM Tag

	A. All records		**B. 2010-2022**		**C. Conserv. Sample**	
	Frequency	%	Frequency	%	Frequency	%
adaptive trial	36	0.00	36	0.00	11	0.01
clinical conference	7,045	0.02	1,724	0.01	12	0.01
clinical study	5,053	0.02	5,050	0.04	641	0.42
clinical trial	498,722	1.53	79,029	0.55	13,969	9.19
clinical trial protocol	9,542	0.03	9,542	0.07	124	0.08
clinical trial, phase 1	23,704	0.07	14,139	0.10	10,951	7.20
clinical trial, phase 2	37,623	0.12	22,749	0.16	17,166	11.29
clinical trial, phase 3	20,597	0.06	15,647	0.11	11,112	7.31
clinical trial, phase 4	2,251	0.01	1,790	0.01	1,250	0.82
comparative study	1,752,380	5.37	424,681	2.97	22,521	14.82
controlled clinical trial	88,132	0.27	13,839	0.10	2,181	1.43
equivalence trial	1,047	0.00	1,047	0.01	505	0.33
evaluation study	243,290	0.75	123,897	0.87	998	0.66
observational study	127,461	0.39	127,370	0.89	5,996	3.94
pragmatic clinical trial	2,112	0.01	2,112	0.01	253	0.17
randomized controlled trial	549,711	1.69	284,774	1.99	79,907	53.57
twin study	9,030	0.03	4,964	0.03	14	0.01
validation study	101,817	0.31	62,261	0.44	495	0.33
	N=32,621,474		N=14,316,494		N=151,997	

Notes: Table A.1 reports the frequency and percentage of records indexed in PubMed that have been categorized by the NLM as falling into each of 18 categories. These categories are selected for their potential to contain a publication satisfying the restriction enumerated in Definition 2.1.

ClinicalTrials.gov (NCT); European Union Drug Regulating Authorities Clinical Trials Database (EUDRACT); International Traditional Medicine Clinical Trial Registry (ISRCTN); Australian New Zealand Clinical Trials Registry (ACTRN).

We note the potential to overcount records using these searches. We focus on instances where each trial identifier prefix is followed by numbers, letters, or punctuation (e.g., NCT12345, ISRCTN: 12345). However, we may collect records that include these characters in other settings (e.g., the world dist**inct**). Table A.2 reports the frequency of each registry identifier flagged across all records in PubMed and subset to those published after 1 January 2010. Figure 1 plots the number of trials with any registry identifier over time. Reassuringly, we observe registry identifiers in PubMed data beginning around 2010, as registration mandates were implemented, and find a steady, smooth increase over time.

A.2.3 Keywords. We flag records that include any keyword likely to indicate that the record reports the results of a clinical trial in their abstract text. We selected these keywords after several reviewing roughly 200 abstracts flagged through the pubtype process, described above.

TABLE A.2. Composition of PubMed by Trial Registry Identifiers in Abstract Text

	A. All records		**B. 2010-present**		**C. Conserv. Sample**	
	Frequency	%	Frequency	%	Frequency	%
NCT	97,923	0.30	94,279	0.66	30,656	20.17
EUDRACT	2,835	0.0	2,816	0.00	1,564	1.03
ISRCTN	12,261	0.03	11,325	0.07	1,275	0.84
ACTRN	5,818	0.01	5,656	0.03	553	0.36
	N=32,621,474		N=14,316,494		N=151,997	

Notes: Table A.2 reports the frequency and percentage of records indexed in PubMed that include a string associated with a clinical trial registry identifier in their abstract text. That is, the first row counts the number and percentage of publications that contain the string "nct."

TABLE A.3. Composition of PubMed by Keywords in Abstract Text

	A. All records		**B. 2010-present**		**C. Conserv. Sample**	
	Frequency	%	Frequency	%	Frequency	%
randomized	550,165	1.69	366,378	2.56	65,637	43.18
controlled trial	107,935	0.33	80,789	0.56	11,535	7.59
control trial	6,049	0.02	4,846	0.03	371	0.24
clinical trial	129,510	0.40	92,035	0.64	17,644	11.61
treatment group	44,627	0.14	27,826	0.19	4,296	2.63
control group	447,901	1.37	286,250	2.00	14,245	9.37
intervention	629,237	1.93	463,373	3.24	10,460	6.88
clinical study	32,652	0.10	18,589	0.13	2,411	1.59
	N=32,621,474		N=14,316,494		N=151,997	

Notes: Table A.3 reports the frequency and percentage of records indexed in PubMed that include each of a collection of keywords in their abstract text. That is, the first row counts the number and percentage of publications that contain the string "randomized."

These keywords are:

> Randomized; Controlled trial; Control trial; Clinical trial; Treatment group; Control group; Intervention; Clinical study.

Table A.3 reports the frequency of each keyword flagged across all records in across all records in PubMed and subset to those published after 1 January 2010.

A.2.4 Intersection. There are 3,925,958 records with at least one of the following attributes: a clinical-trial indicative NLM tag; a clinical trial registry identifier; a clinical-trial indicative keyword in the abstract text. The three categories overlap. Table A.4 records the size of the overlap of each category. In the main text, we restrict attention to this sample for the years 2010-2022. This sample includes 1,821,429 publications.

A.2.5 Novel Census. We add columns to Tables A.1 to A.4 that document the frequency of each flag in the data constructed for this paper (we use the conservative sample). Several facts are worth noting. First,

TABLE A.4. Overlap of Publication Attributes

Records with any:	that have any:	**A. All records**	**B. 2010-present**	**C. Conserv. Sample**
Registry ID		116,564	111,837	33,040
	NLM tag	81,894	78,366	26,885
	Keyword	73,164	70,414	19,103
NLM tag		2,858,924	1,804,081	124,260
	Registry ID	81,894	3,528	26,885
	Keyword	515,755	220,681	67,153
Keyword		1,564,659	1,044,461	91,349
	Registry ID	73,164	70,414	19,103
	NLM tag	515,755	295,074	67,153

Notes: Table A.4 reports the size of the intersection between three sets of records indexed in PubMed. These categories are defined by the properties: Possession of a clinical-trial indicative NLM tag, inclusion of a clinical trial registry identifier in abstract text, and inclusion of a clinical-trial indicative keyword in abstract text. These properties are defined in Appendices A.2.1 to A.2.3, respectively.

Table A.2 suggests that—between 2010 and 2022—less than 25 percent of records identified as reporting the results of a clinical trial reported a registry identifier in their abstract. Although, in principle, registry identifiers may be reported in publication full-texts, rather than abstracts, the low frequency of identifiers in our sample suggests non-compliance with requirements, consistent with patterns in ClinicalTrials.gov documented in DeVito et al. (2020). Second, Table A.1 sheds light on the differences in trial composition that we observe in Figure 2.1: NLM tags that, ostensibly, capture clinical trials have little overlap with one another and do not fully capture the records in our sample. Of the roughly 150,000 records in our census, approximately 100,000 could be identified using NLM tags for "clinical trial" (and its variants) and "randomized controlled trial." Observe, however, that use of such flags captures a much larger number of records than one would intend to include. Finally, Table A.4 suggests that although various text-based flags that may indicate a record is a clinical trial, in the sense of Definition 2.1, are correlated, there is less overlap than one might expect.

A.3 Determining the Set of Leading Medical Journals

Angrist et al. (2020) propose a strategy to identify the leading journals in a field, which we adapt to this context. In medicine, the "trunk journals," per Angrist et al. (2020), are the Journal of the American Medical Association and the New England Journal of Medicine. We collect all instances in which papers published in these journals cite other records in PubMed, between 2010 and 2022. In Angrist et al. (2020), the authors select the fifty journals receiving the most cites from these trunk journals. As medicine contains a large number of subfields, we instead collect a list of journals that receive at least 100 citations from a trunk journal over this time period, which yields a list of 84 journals.

In our data, papers published in trunk journals cite records in 3,136 unique journals.[49] The median journal is cited twice by these two trunk journals over our time period. The most frequently cited journals are, perhaps unsurprisingly, the trunk journals themselves: JAMA receives 3,143 citations, and the New England Journal of Medicine receives 4,302 citations. The next ten journals, by frequency of citations, are: Lancet (1,457), Circulation (789), Journal of Clinical Oncology (734), Nature (660), Blood (651), PLoS One (645), Journal of the American College of Cardiology (593), Clinical Infectious Diseases (589), BMJ (586), and Annals of Internal Medicine (586).

A.4 The Web of Science

We collect supplementary information about papers in our sample from Clarivate Analytics' Web of Science.[50] Web of Science data allow us to examine more detailed information about each author in our sample. While PubMed includes some details on authors, including limited institutional affiliation and address information, the Web of Science provides standardized author addresses for roughly 75 percent of observations in our sample.

For each paper in our sample, we extract countries from available mailing addresses for first- and last-listed authors. First- and last-authors have the same listed country for 85 percent of clinical trial records and 91 percent of non-clinical trial records. Thus, we impute to each record the country associated with the first-listed author. In Figure 3.4, we plot geographic trends over time.

We use the Web of Science to collect details about author career and publication histories, which allow us to investigate patterns in the production of medical research. The Web of Science assigns to authors a unique, persistent identifier (*DAIS ID*), which allows us to collect information on authors' output before and after the publication focal records in our sample. To our knowledge, there have been no large-scale efforts to validate the quality of these researcher identifiers. On inspection, it appears as though the identifiers perform best, perhaps unsurprisingly, for researchers with distinctive names. When researchers have names that are more frequent in the population, our spot-checking (against online publication records) suggests that there are more likely to be errors. Nonetheless, tracking researchers using these identifiers is standard in this literature, as there are no obviously superior alternatives. Especially given that researcher productivity over time is not a centerpiece of our analysis, we do not undertake additional validation exercises.

[49] We use unique entries in the field *medlineta*—which correspond to distinct MEDLINE abbreviations for journal names—to identify distinct journals. See `https://www.nlm.nih.gov/bsd/serfile_addedinfo.html` for details on name abbreviations in this database. We confirm that the count of journals is identical if we, instead, use journals identified by the National Library of Medicine's unique journal identification number.

[50] We use a copy of the Web of Science licensed to Stanford University.

APPENDIX B. CLINICALTRIALS.GOV

ClinicalTrials.gov is a registry of clinical trials, run by the United States National Library of Medicine. As of 1 August 2024, ClinicalTrials.gov lists more than 500,000 studies in all 50 states and in 222 countries.[51] ClinicalTrials.gov was available for study registration beginning 29 February 2000. At this time, there were 1,255 registered studies.[52] By the beginning of 2006, there were 24,822 studies and by the beginning of 2012, 118,020. In January 2024, there were 477,227. The geographic composition of these studies has shifted considerably over time, as have reporting patterns from various firms and universities.

In this Appendix, we use data from ClinicalTrials.gov to illustrate its limitations as a source for the type of high-quality records necessary for studies of productivity. Given the parallels between this paper and a long literature that studies the usefulness of patent data for studies of innovation, it is worth highlighting the important, perhaps obvious, distinction between ClinicalTrials.gov as a government database and records collected by the U.S. Patent and Trademark Office (USPTO). Records of U.S. patents are collected by the USPTO itself. Disclosure is widely considered a pivotal component of the patent system: in exchange for a short-term quasi-monopoly, information on patented inventions is placed in the public domain. Although ClinicalTrials.gov is described in similarly forceful terms in the laws that authorize its creation, disclosure has been—and continues to remain—effectively voluntary, in the sense that legal tools have never been used by a federal agency to compel the registration of a study or disclosure of its findings.

Institutional details suggest that the "voluntariness" of this reporting has shifted over time, in ways that likely affected sponsors' propensity to register.[53] A coalition of medical journal editors began encouraging registration in 2005 (Laine et al., 2007), registration requirements were imposed for certain FDA-regulated trials in 2008, and the FDA began threatening to impose fines on noncompliant sponsors in 2021. In Table B.1, we collect a list of five changes in the past 25 years that changed sponsors' requirements for trial registration.

We inspect a version of the Aggregate Analysis of ClinicalTrials.gov (AACT) data, a publicly available relational database that contains information on all studies registered on ClinicalTrials.gov, downloaded on 2023 May 31. Of interest, for us, is the question of whether ClinicalTrials.gov can provide an accurate count of the total quantity of clinical trials and, if so, at what date such an accurate count might be available. Figure B.1 displays the results of our investigation.

[51] See https://clinicaltrials.gov/about-site/trends-charts.

[52] In 1988, in response to public pressure surrounding perceived lags in the development of therapeutics and preventives to treat HIV/AIDS, U.S. Congress passed the Health Omnibus Programs Extension Act (Public Law 100-607), which required the creation of a database of AIDS Clinical Trials Information Services (ACTIS). Publicly-funded clinical trials that were associated with these previous data registration efforts were included in ClinicalTrials.gov at its inception.

[53] In guidance to researchers on the use of ClinicalTrials.gov, Tse et al. (2018) collect a list of the ten challenges associated with using these data, chief among which is the fact that changes in medical journal reporting requirements and federal law make it extremely difficult to determine if increases in the size of the database are related to actual increases in the amount of scientific research.

TABLE B.1. Policy Changes Affecting the Composition of ClinicalTrial.gov

Policy (Effective Date)	Registration Requirement
Food and Drugs Administration Modernization Act (21 November 1997; ClinicalTrials.gov registry available starting on 29 February 2000)	Clinical trials of investigational new drugs for serious or life threatening conditions and diseases
International Committee of Medical Journal Editors (1 July 2005 for newly initiated trials and 13 September 2005 for ongoing trials)	All clinical trials as a condition for consideration of publication of results
Food and Drugs Administration Amendments Act (FDAAA) (27 September 2007)	Non-phase 1 clinical trials of FDA regulated drug and biological products, and non-feasibility trials of FDA regulated device products
Final rule implementing FDAAA in Title 42, Part 11 of the Code of Federal Regulations (42 CFR Part 11) (18 January 2017)	Non-phase 1 clinical trials of FDA regulated drug and biological products, and non-feasibility trials of FDA regulated device products

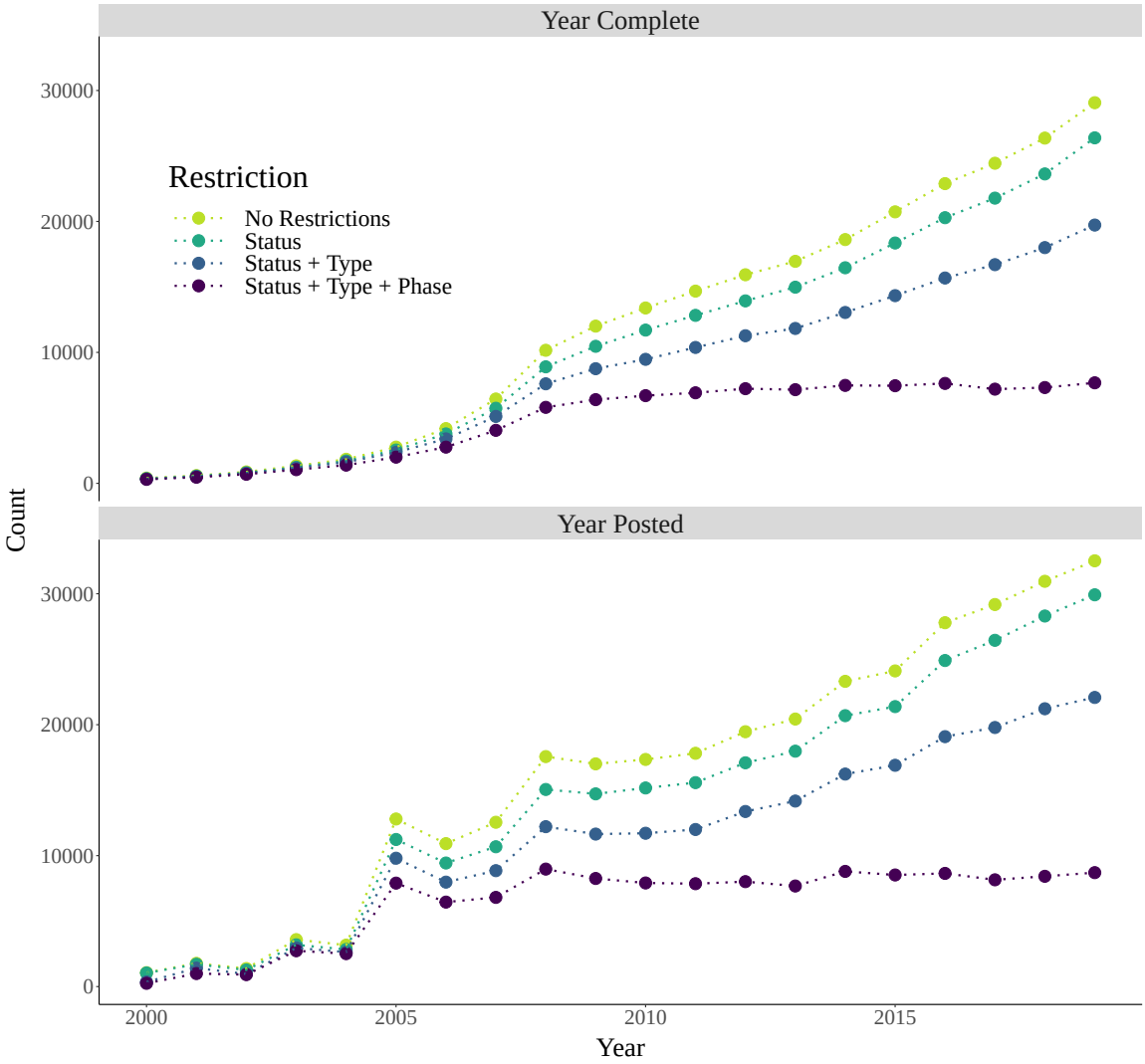
Notes: Table B.1 collects a list of five policy changes in the past 25 years that changed sponsors' requirements for trial registration. This table is adapted from Tse et al. (2018).

We plot counts of records, selected in various ways, against two x-axes. In the top panel of Figure B.1, we plot counts associated with the year in which clinical trials are marked complete.[54] In the bottom panel, we plot counts based on the year in which the study was posted on ClinicalTrials.gov. Measures of trial completion are informative, insofar as they capture the date on which information production is finished. However, trials vary greatly in length—the key observation in Budish et al. (2015)—and as a result represent investments "sunk" at very different earlier points in time. Measures of trial posting are analogous, in some ways, to publication, in that they capture decisions to disclose information. Even without considering the specific series in Figure B.1, observe that the bottom panel includes a variety of sharp changes in the count of records. These correspond to several of the policy changes in Table B.1. We see a sharp change in all four series in 2005, when medical journal policies were changed, a sharp change in 2008, when the FDAAA took effect, and a sharp (albeit smaller) change in 2016, in the lead-up to the effective dates of the final two policy changes in Table B.1. We observe a smaller, but discernible, change in the slope of each series in 2008 when measures of trial completion are used. We draw attention to these differences, in part, because an important challenge for the use of these records in research is determining the appropriate measure of time.

Turning to the content of these series, we begin by examining the total number of records indexed in ClinicalTrials.gov over time (light green). We define a record as a unique "national clinical trial identifier," although Tse et al. (2018) note that one study may be assigned multiple NCTs. There is no method, to our knowledge, that further distinguishes between records. In both panels, this series begins at zero in the year

[54]These completion dates are somewhat imprecise. If a trial is registered, but no results are reported, the completion date is an estimate, based on the projected completion date in the study's protocol. If the trial is registered and has associated results, the completion date is the actual date on which data collection ended.

FIGURE B.1. Number of Entries in Clinical Trials.gov, by Sample Restrictions



Notes: Figure D.1 displays the counts of the number of records in ClinicalTrials.gov that satisfy various sample restrictions. The x-axis of the top panel is the year in which the clinical trial was completed. The x-axis of the bottom panel is the year in which the clinical trial was registered on ClinicalTrials.gov.

2000 and increases sharply, to roughly 30,000 records by the end of our data, 2022. On inspection, it is clear that many of the 30,000 records included in the 2022 bin are not clinical trials. Among those that are clinical trials, many do not study medicines, but rather randomize exposure to interventions involving psychological treatments and exercise.

We impose three sets of restrictions, which aim to bring the data closer to a consistent sample of clinical trials studying the effects of a drug. We select these restrictions based on examination of various fields in ClinicalTrials.gov and with the aim of providing an illustration of our concerns with the use of this dataset

"off the shelf." There are alternative ways of constructing a sample of records that may be necessary to yield relevant data for different research questions.

We begin by dropping a set of records for which the study "status" variable in the database suggests that the study was either never started or was terminated. Specifically, we drop 40,622 records for which the status is listed as "Withdrawn," "Suspended," or "Terminated." This status restriction is plotted in teal and has little impact on the trends in either panel.

Next, we observe that not all records in ClinicalTrials.gov are interventional studies. We drop any study with a type other than "interventional." This removes an additional 98,138 observations from our data. This status and type restriction is plotted in blue. We observe that the sharply increasing trend in the later years of our data flattens out slightly. That is—the growth in the quantity of records is driven, in part, by growth in the number of registered non-interventional studies.

We implement a third restriction. We observe that clinical trial "phases" are a hallmark of drug trials: medical devices, for example, are approved via a "stage" process. Any record for which the trial phase is either missing or "Not Applicable" is then less likely to be a drug. We drop these records. This removes 151,541 observations from our data. We plot the resulting series in purple. When all three restrictions are imposed, we find steady increases in the quantity of trials leading up to 2008 and a stable series from 2008 forward. Reassuringly, this trend mirrors that documented in Figure 3.1. These plots suggest a level difference of roughly 2,000 observations, where Figure 3.1 is higher. This is consistent with the idea that ClinicalTrials.gov is *not* a global registry and, as Zarin et al. (2017) notes, many sponsors continue not to register studies in ClinicalTrials.gov

We emphasize that it is not obvious—in the absence of corroborating evidence, similar to that in Figure 3.1—that the version of this series with all three restrictions in Figure B.1 provides an accurate measure of trends in research, especially given the sharp changes in the early part of the series and the somewhat haphazard way that we identified a set of sample restrictions.

A series of recent papers—including DeVito et al. (2020) and DeVito and Goldacre (2021)—document limitations with the quality of particular data elements, which further suggests the difficulty of using this database as a standalone source of estimates.

Appendix C. Prompt Design, Fine-Tuning, and Performance Assessment

In this appendix, we detail the construction of the sample of clinical trials studied in Section 3. We begin in Appendix C.1 by describing the process we used to hand-label a sample of publications according to whether they satisfy the restrictions enumerated in Definition 2.1. In Appendix C.2, we document poor performance of several approaches for classifying publications according to whether they belong in our sample based on standard machine learning methods. Appendix C.3 overviews the process of constructing prompts that perform well when used to query proprietary language models.

C.1 Hand-Labeling

A standard concern when using large language models is the quality of the output. We hand-label a set of 3,000 abstracts, selected randomly from our candidate set of records, which allow us to quantify the performance of our classification procedure.

We devise an interface for abstract labeling, which allows us to review records quickly. Figure C.1 provides an example of this interface. A labeller reads in a set of publication records and is shown one PubMed identifier (PMID) and one abstract at a time. The labeller can select that the abstract satisfies our sample criteria, or else can indicate a reason that the abstract should be excluded. The buttons for exclusion mirror the inclusion and exclusion criteria in Definition 2.1, using short-hand convenient for labellers. We exclude records that report, only, the protocol for a clinical trial without results ("protocol"), that report the results of an observational ("observational") or retrospective ("retrospective") study, that report the results of a study involving animals ("animal"), that do not include human subjects ("no human subjects"),[55] that do not study a drug ("no drug"), that report the results of a specific clinical case or patient experience ("case report"), or that summarize the findings of existing work ("meta-analysis"). By our own estimates, the labelling interface increased the speed of labelling by a factor of seven. Each record was reviewed, and labelled, twice.

The hand-labeled data are split into three subsets—validation, training, and testing—based on their eventual use. Splits are assigned randomly. We assign 1000 records to a test set, 1000 to a validation set, and 1082 to a training set. Of these records, we assign labels only to records that have abstracts. Our final training data includes 1082 records, our validation dataset includes 1000 records, and our test set includes 993 records. The loss of seven records in our test set reflects a coding error. We add additional records, randomly selected, to the training and validation datasets when abstracts are missing. We do not add extra records to the testing dataset.

C.2 Benchmark Comparison to Standard Machine Learning Methods

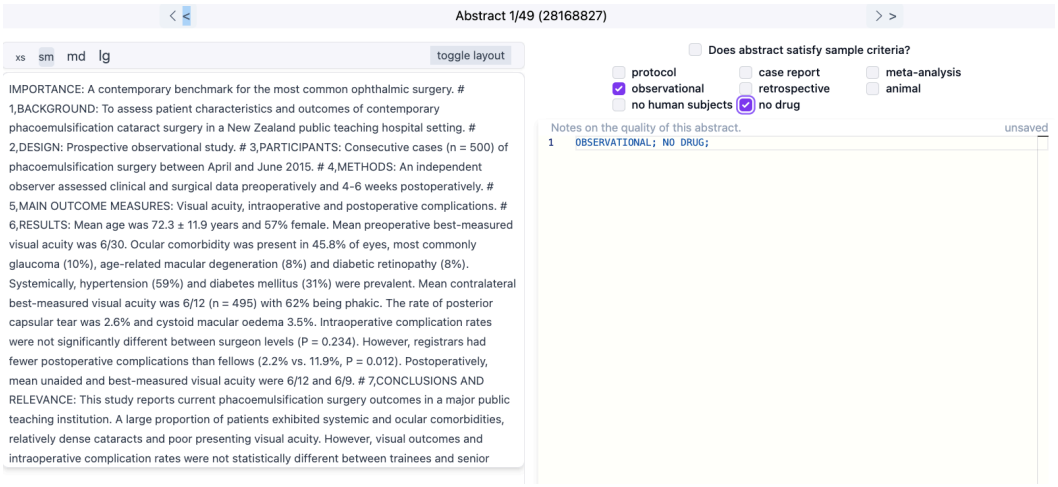
We measure the performance of a variety of standard machine learning algorithms for classifying publications according to whether they meet our sample restrictions. Models are trained in the training and validation samples and tested in the testing sample.

We compute two embeddings for each abstract in our hand-labeled sample. First, we compute Term Frequency, Inverse Document Frequency (TF-IDF) embeddings based on the corpus of abstracts in the hand-labeled sample. Second, we extract embeddings associated with the abstract of each publication in the hand-labeled sample with the SENTENCETRANSFORMERS language model (Reimers and Gurevych, 2019). The SENTENCETRANSFORMERS language model is a standard source for embeddings, and is a modification of the BERT language model (Devlin et al., 2019).

We consider six classes of prediction algorithms: logistic ridge regression, logistic lasso regression, Support Vector Machines, Random Forest Regression, Boosted Regression, and Convolutional neural networks. Each

[55] In practice, "animal" and "no human subjects" catch different sets of records. We flag records that include *any* animal with "animal." Many studies in our sample of candidate records occur *in vitro*—in test tubes or other laboratory-based settings. There, no living subjects of any kind are enrolled and the "no human subjects" button is used.

FIGURE C.1. Abstract Labeling Interface



Notes: Figure C.1 displays a screenshot of the application we developed to hand-label abstracts. Each abstract was labelled according to whether it satisfied the sample criteria enumerated in Definition 2.1. If a publication did not satisfy our sample restrictions, we stipulated a reason. The user interface was developed with the "Meerkat" Python package available at the link: https://github.com/HazyResearch/meerkat.

model is obtained from the "sklearn" Python package. The hyper-parameters of each model are chosen with 10-fold cross-validation implemented in the training and validation samples. A balanced loss is used to measure the performance of each hyper-parameter.

Figure 2.2 displays receiver operating characteristics for the each class of models estimated with both types of embedding. Error rates are estimated in the testing data. The performance of the receiver operating characteristic for the "ensemble" model whose performance is displayed in Figure 2.4 is given in purple. The fine-tuned large language model developed in this paper substantially out-paces the performance of standard machine learning methods.

C.3 Prompt Design and Error Analysis

We extract weak labels for 64,000 randomly selected abstracts with two proprietary large language models: OpenAI's GPT-3.5 and GPT-4 (Nori et al., 2023; Bubeck et al., 2023). As discussed in the main text, we identify three general prompt formats, which differ both in the amount of detail provided about our classification task and in the structure of the requested model completion. We refer to these prompts as Prompt 1.0, Prompt 2.0, and Prompt 3.0, respectively. The text of these prompts are displayed in Figure E.1, Figure E.5, and Figure E.7, respectively.

We test each of these three prompts in our 1000-record hand-labelled validation dataset, using both GPT-3.5 and GPT-4. We conduct a detailed error analysis, reported in Table C.1. To make this concrete, suppose that GPT-3.5, given Prompt 1.0 and an abstract that reports the results of a meta-analysis, erroneously classifies the record as clinical trial, per our definition. In Table C.1, we flag this error under Prompt 1 for GPT-3.5,

as an error of type "meta-analysis." Within meta-analysis, we assign the error to a sub-type. If the abstract explicitly includes terms such as "meta-analysis," "literature search," or "literature review," we categorize this as an explicit error. If the abstract references that the publication is summarizing existing studies, or searching a database for records and collecting their findings, we categorize this as an implicit error. We devise error types based on our inclusion/exclusion criteria, in Definition 2.1, and select sub-types based on common categories of errors.

We revise each class of prompt based on these findings. For Prompt 1.0, the simplest true/false prompt, we consider three variants. We refer to these revised prompts as Prompts 1.1, 1.2, and 1.3. The text of these prompts is displayed in Figures E.2 to E.4. For Prompts 2 and 3, we consider one variant each. We refer to these revised prompts as Prompts 2.1 and 3.1. The text of these prompts is displayed in Figures E.6 and E.8. These changes reflect the differences in performance catalogued in Table C.1.

TABLE C.1. Baseline Prompt Error Analysis

Model	GPT-3			GPT-4		
	Prompt			Prompt		
Error Type	1.0	2.0	3.0	1.0	2.0	3.0
No Drug						
vitamin/supplement	12	12	9	15	4	15
surgical/medical procedure/diagnostic	40	27	29	35	21	29
abstract not specific about name of medicine	3	4	4	6	5	6
food/beverages	8	3	6	13	1	12
surgical/medical device	8	4	5	7	2	6
surgical/medical imaging	1	1	1	2	0	2
behavioral/physical therapy/exercise	40	22	36	25	8	13
misses the mention of a drug	25	10	5	2	1	0
surgical/medical material (e.g., resin/dental filling)	1	0	1	3	1	1
non-medical pollution/chemical/drug	1	1	1	1	2	1
other	1	0	1	2	1	2
Meta-Analysis						
explicit mention of meta-analysis/literature search/literature review	13	5	13	8	0	6
mention of database searched, explicit mention of meta-analysis	3	3	3	3	0	2
summary existing studies, without reference to data search / meta-analysis	6	7	4	3	4	3
Retrospective						
re-analysis of previously collected data without explicitly mentioning “retrospective”	8	2	6	6	5	4
explicitly described as “retrospective” or “retrospective analysis”	6	1	6	0	0	0
miscategorizes a study conducted in the past as “retrospective”	0	1	0	0	0	0
other						
Observational						
no active intervention described; does not explicitly say “observational”	10	9	11	6	0	4

no active intervention described; explicitly says “observational”	14	13	13	13	2	11
misses reference to active intervention	1	1	1	0	0	0
Protocol						
no results reported about current study; explicitly says “protocol”	10	9	11	6	0	4
no results reported about current study; does not explicitly say “protocol”	14	13	13	13	2	11
study is based on a simulation, not real world data	1	1	1	0	0	0
No Human Subjects						
only references to laboratory experiments, cells drawn from humans	0	1	0	0	0	0
Animal						
study conducted on animals	5	10	6	2	3	2
Other						
overly literal interpretation of inclusion criteria	0	2	30	0	0	2
studying outcomes/objects unrelated to a randomized trial	1	1	1	1	1	1

Notes: Table C.1 categorizes errors documented in the first round of prompt iteration, using Prompts of sub-type 0. See Table 2.1 for details on model performance. For each model error—an instance in which a model returned a classification that deviated from hand-labelled data—we examined the associated record. We categorized errors by Type and Sub-Type. Types are drawn from the exclusion restrictions in Definition 2.1. Sub-types describe consistent characteristics of publications that resulted in such an error. We report counts of errors, by model and prompt, of each Type and Sub-type.

C.4 Error Analysis

The Conservative model incorrectly labels 27 papers. We inspect each of these errors. We review each abstract and aim to understand what might have generated an error. In roughly half of cases–13 of 27—there is a clear error. These clear errors include publications that explicitly report the results of observational studies (one error), that re-analyze existing data (eight errors), and that report literature reviews (one error). This set also includes clinical trials that satisfy all criteria in Definition 2.1, except the treatment being studied is not a drug (three errors). In 14 cases, however, errors are associated with records that are difficult for human labellers to categorize, either because the content of the publication does not fit neatly into the inclusion and exclusion criteria implied by Definition 2.1 or because the publication is written in a way that makes it difficult to determine details of the study. PubMed record 27880726 is illustrative. This publication studies the effect of chrysophanic acid (CA) on benign prostatic hyperplasia. The abstract is unclear about whether this treatment is studied in human subjects or in an animal model–though the publication title and full-text make clear that this was conducted in a sample of rats. Based on the abstract alone, however, nothing suggests that this is an animal study. It is excluded from both the conservative and moderate model-generated samples, but a human-labeller flagged it as satisfying our criteria.

Appendix D. Additional Figures and Further Analyses

This appendix collects additional figures and analyses.

D.1 Robustness of Trends

Figure D.1 displays series analogous to the series displayed in Figure 3.1, constructed with the Moderate and Liberal sample restrictions. The results are very similar.

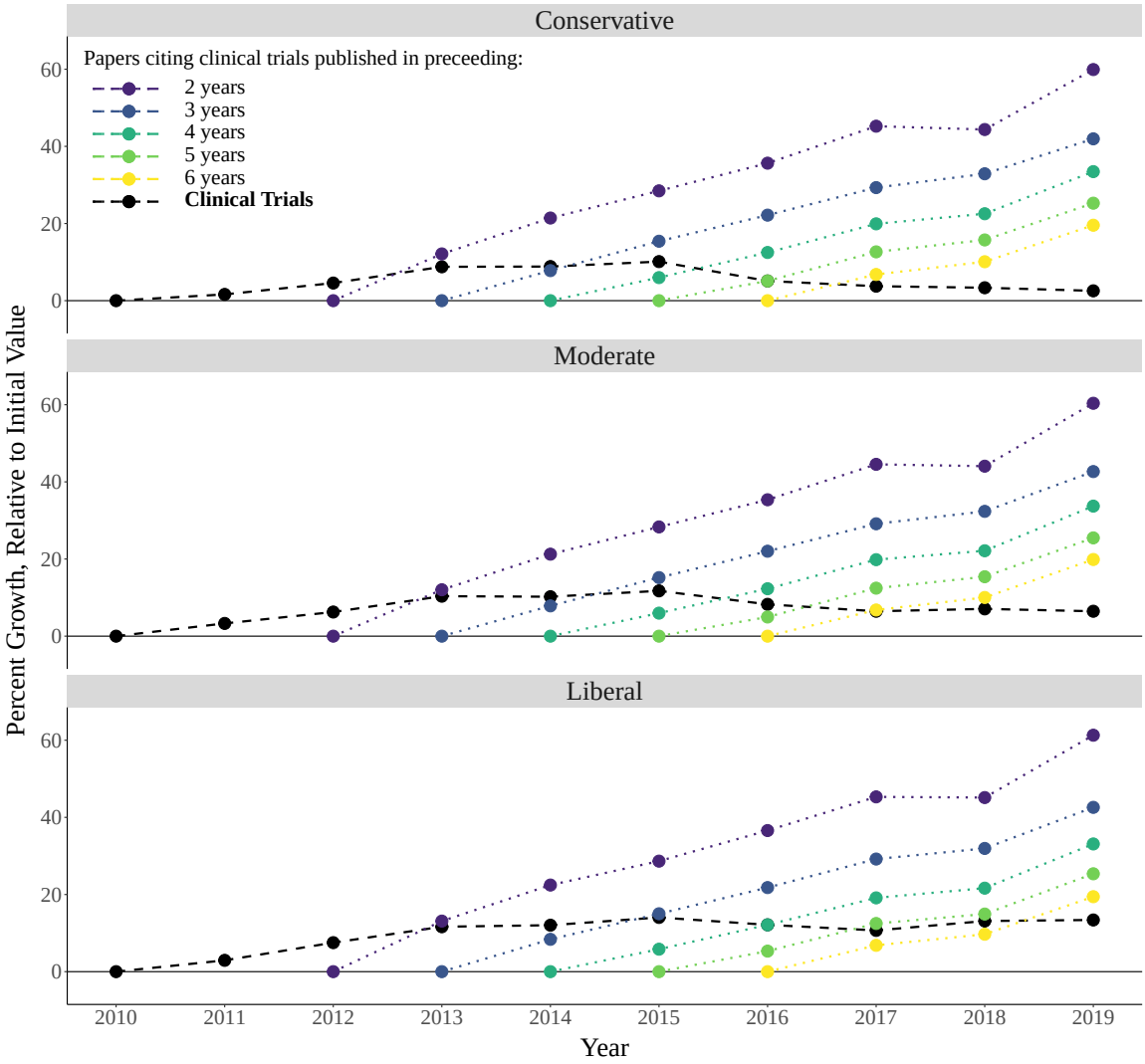
One concern with the comparison made in Figure 3.1 is that researchers may cite a growing number of publications over time. That is, more papers may cite clinical trials over time because citations are becoming, in some sense, "cheaper." The relevant clinical literature, then, may not be any larger. To alleviate such concerns, Figure D.2 reports counts, where we scale each citation assigned to a paper by the total number of citations in the originating paper. That is, a paper that receives two citations, each from a paper that cites two papers, will have a weighted citation count of 1.

Figure D.3 displays series analogous to Figure 3.2 with the Moderate and Liberal sample restrictions.

Figure D.4 gives a displays analogous Figure 3.3 for the distribution of citations received by clinical trials in each calendar year originating from all publications. Figure D.5 gives a displays analogous to Figure 3.5 for the distribution of citations received by publications citing clinical trials (three year sample) originating from all publications.

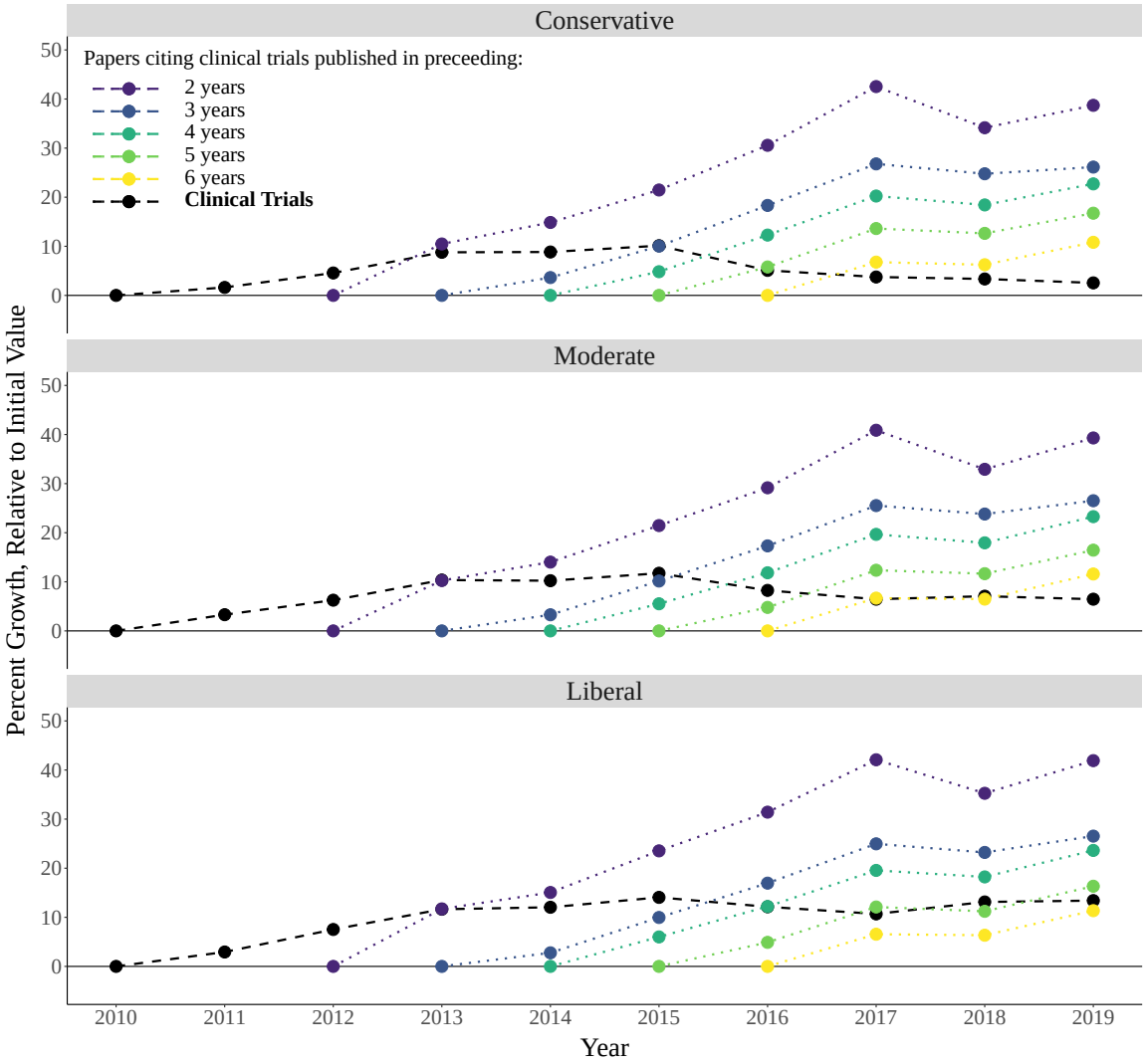
Figures D.6 and D.7 display heat maps analogous to Figure 3.3 and Figure D.4, with the Moderate and Liberal sample restrictions and both 3-Year and 5-Year citation counts. In each case, the results are very similar across specifications.

FIGURE D.1. Growth in Medical Research, by Sample Stringency



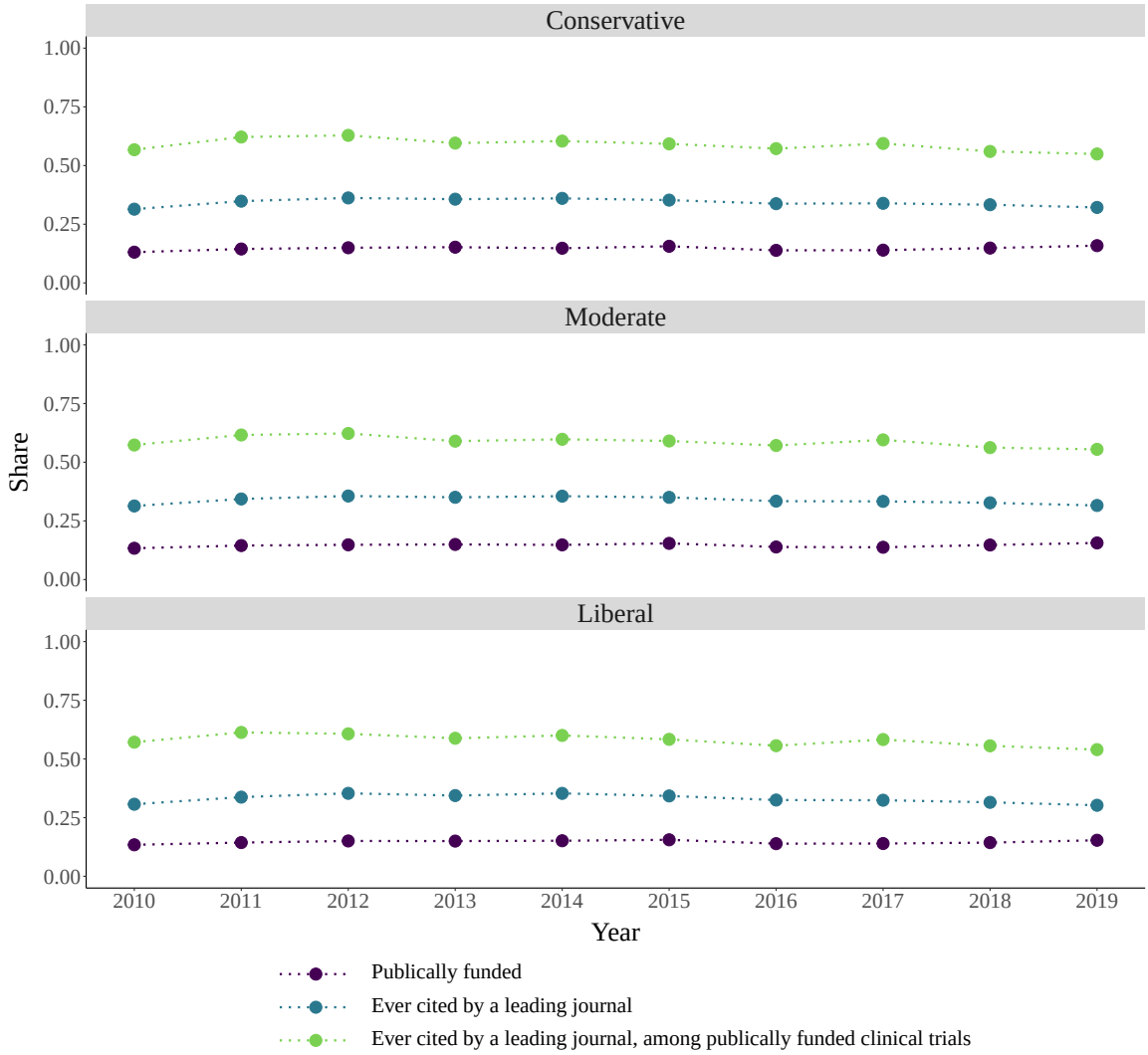
Notes: Figure D.1 displays measurements of the number of clinical trials, and papers that cite clinical trials, published in each calendar year, in all three cuts of our data—which we term "conservative," "moderate," and "liberal." Each series is reported in terms of the percent change relative to its initial value. To address truncation, we report the number of publications that cite clinical trials published in the preceding t years for each t between 2 and 6.

FIGURE D.2. Growth in Cost-of-Citation Weighted Medical Research



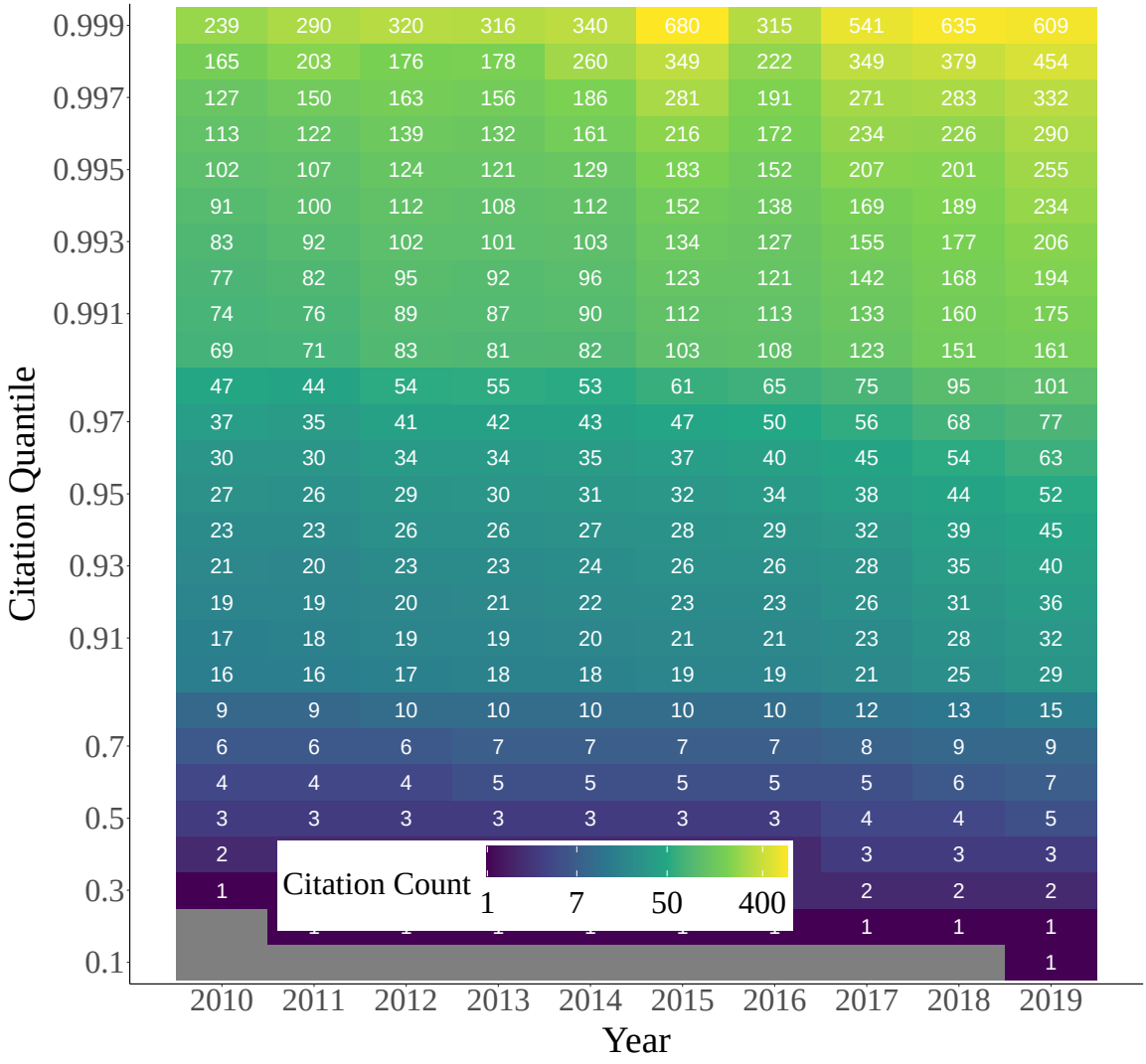
Notes: Figure D.2 displays measurements of the number of clinical trials, and papers that cite clinical trials, published in each calendar year, in all three cuts of our data—which we term "conservative," "moderate," and "liberal." Here, we report the number of scientific papers using a citation-weighted metric. In particular, we scale each citation assigned to a paper by the total number of citations in the originating paper. Each series is reported in terms of the percent change relative to its initial value. To address truncation, we report the number of publications that cite clinical trials published in the preceding t years for each t between 2 and 6.

FIGURE D.3. Quality Measures, by Sample Stringency



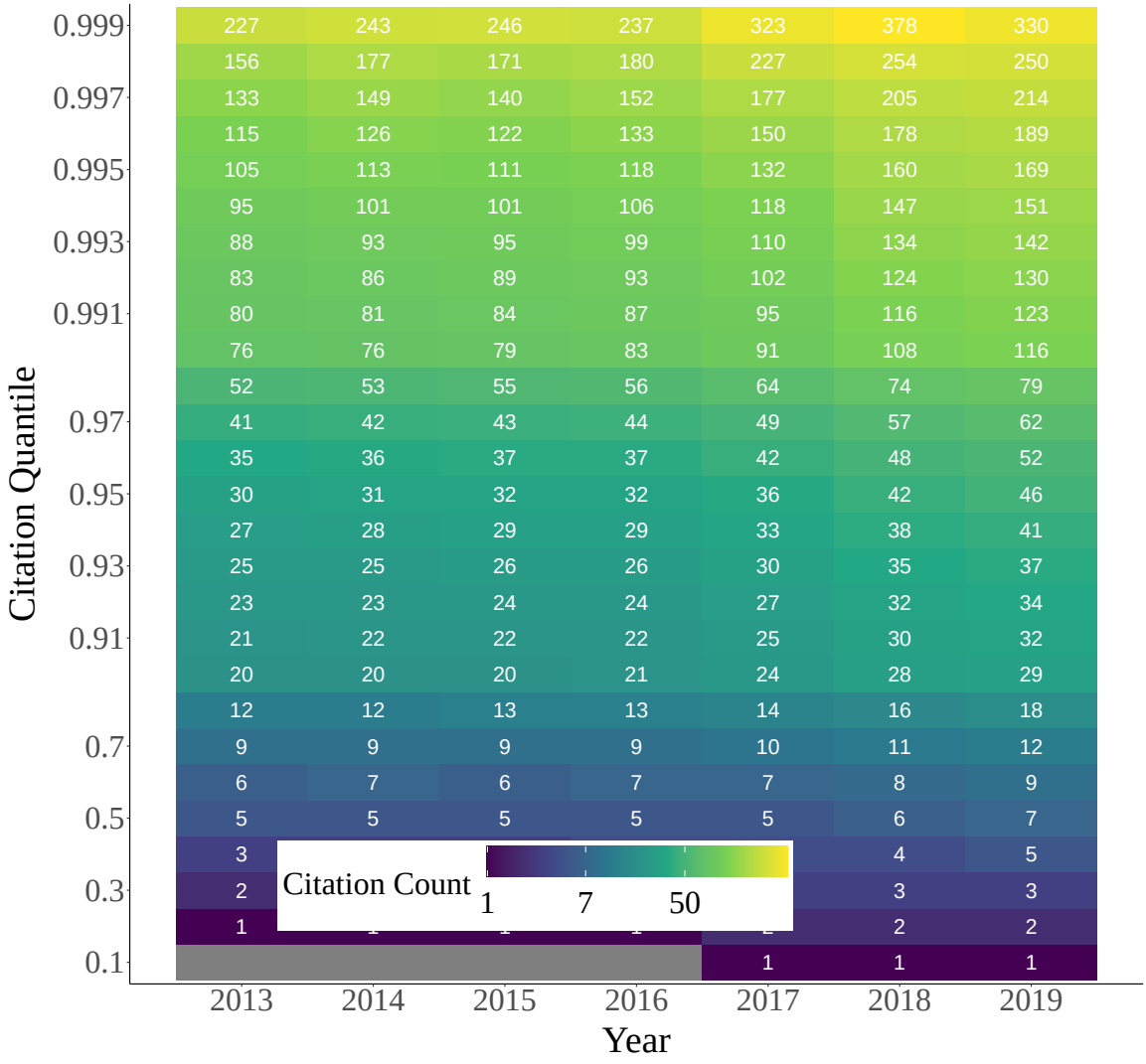
Notes: Figure D.3 illustrates the stability of the heterogeneity in the composition and quality of published clinical trials between 2010 and 2019 in all three cuts of our data—which we term "conservative," "moderate," and "liberal." See Figure 2.4 for the true positive and false positive rates of these models. The figure displays three time series: the proportion of clinical trials that are publicly funded, the proportion of clinical trials that are ever cited by a leading journal, and the proportion of publicly funded clinical trials that are ever cited by a leading journal.

FIGURE D.4. Citation Distribution Across Time, All Publications to Clinical Trials



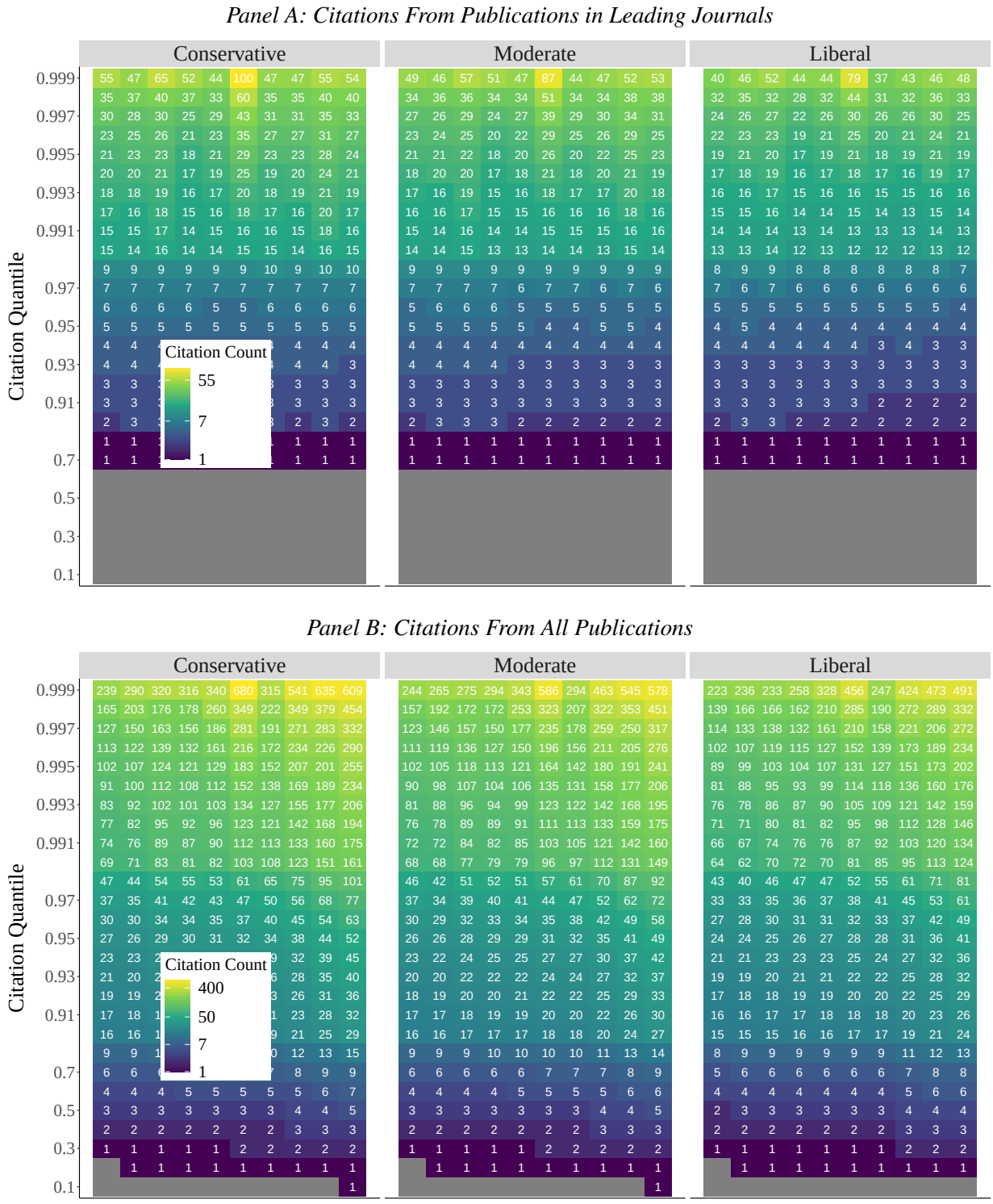
Notes: Figure D.4 displays a heat map measuring the distribution of citations received by clinical trials in each calendar year from all publications. The y-axis has been stretched to elongate the right-tail of the citation distribution. Colors are displayed in a log scale. The sample of published clinical trials is constructed with the conservative model.

FIGURE D.5. Citation Distribution Across Time, All Publications to Citing Publications



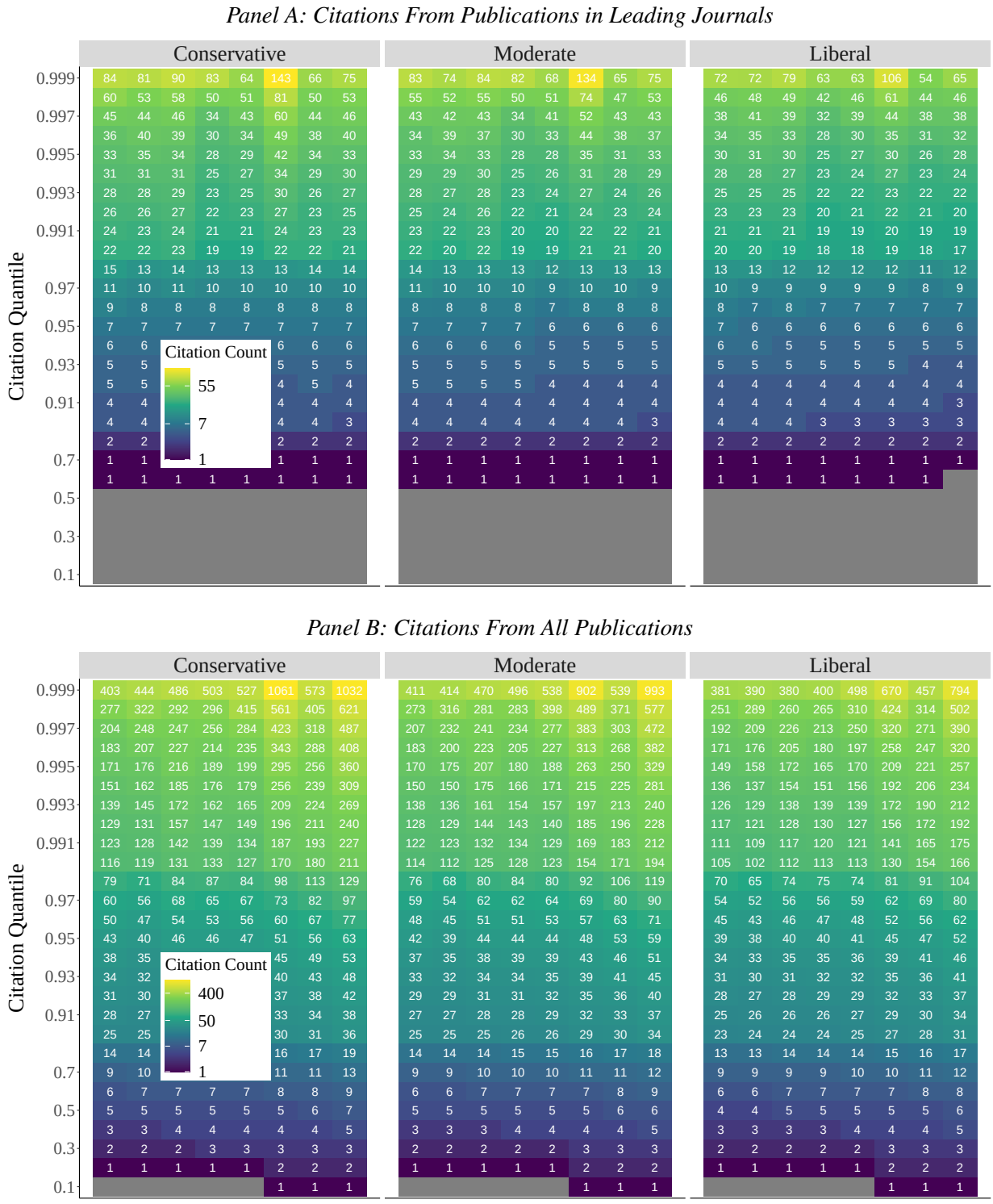
Notes: Figure D.4 displays a heat map measuring the distribution of citations received by papers that cite clinical trials, in each calendar year, from all publications. The y-axis has been stretched to elongate the right-tail of the citation distribution. Colors are displayed in a log scale. The sample of published clinical trials is constructed with the conservative model. The sample of papers that cite clinical trials is constructed using a three-year window.

FIGURE D.6. Distribution of 3-Year Cites, Clinical Trials, by Sample Stringency



Notes: Figure D.6 displays heat maps measuring the distribution of three-year citations received by clinical trials in each calendar year from leading journals and from all publications, respectively. The y-axis has been stretched to elongate the right-tail of the citation distribution. Colors are displayed in a log scale. We consider the sample of clinical trials constructed with the conservative, moderate, and liberal model.

FIGURE D.7. Distribution of 5-Year Cites, Clinical Trials, by Sample Stringency



Notes: Figure D.7 displays heat maps measuring the distribution of five-year citations received by clinical trials in each calendar year from leading journals and from all publications, respectively. The y-axis has been stretched to elongate the right-tail of the citation distribution. Colors are displayed in a log scale. We consider the sample of clinical trials constructed with the conservative, moderate, and liberal model.

D.2 Growth in Citing Papers by Geography

In Section 3.3.2, we document differential trends in the production of publications by type. Specifically, we find essentially stable geographic patterns in the production of clinical trials, whereas we find that large increases in the quantity of papers in our citing sample are driven by research produced outside of the United States. Here, we provide additional details on this geographic fact. In Appendix A.4, we provide details on the data that are used to construct these plots.

There are two especially striking aspects of Figure 3.4: the quantity of citing papers published by researchers in China is high and sharply rising, and the quantity of (pooled) citing papers from what we refer to as the "rest of the world" set is also high and rising. Table D.1 breaks down the publication of citing papers by country. We elect to focus on countries that produced at least 200 citing papers in 2019—a cut that captures 93 percent of the citing papers in our data. That is, nearly all of the papers in the citing sample are produced by this relatively small number of countries.

We highlight several patterns in Table D.1. First—consistent with its presentation as a stand-alone series in Figure 3.4—the quantity of citing papers produced by Chinese researchers, the change in the size of this set, and its growth over our sample period are distinctive. While the United States continues the produce the largest number of papers in the citing sample, the growth of this set ranks last among the countries listed in Table D.1. Whereas there is 11 percent change in the quantity of citing papers produced by United States-based researchers, there is more a more than 200 percent change for Chinese papers.

D.3 Quantifying Meta-Analyses

In Section 3.3.1, we present three methods that allow us to identify "review" publications—publications that summarize the findings of existing studies—in our set of citing papers. Each approach is imperfect (for reasons that mirror the motivation for the technical contribution of this paper). Here, we provide additional details on these three approaches.

D.3.1 Hand-Labeling Meta-Analyses. From the set of publications that cite clinical trials (three-year sample), we drew 100 observations randomly from each publication year. We inspected records in the first publication year (2013) and last (2019). Each record was reviewed twice. During the first round of review, we collected notes about the content of the publication. We standardized these notes and observed that these notes could be coded into seven distinct categories. We reviewed the records a second time and assigned these standardized codes. Table D.2 reports the composition of these 100-record samples.

The three largest categories in both samples are pre-clinical studies, observational studies, and meta-analyses. We estimate a decrease in the proportion of pre-clinical studies and an increase in the proportion of meta-analyses. By contrast, the proportion of observation studies appears stable.

D.3.2 Text Searches for Meta-Analyses. Recall that, in the process of constructing the training data that were used to fine-tune our large language models, we identified many records in PubMed that were meta-analyses,

TABLE D.1. Changes in the Number of Papers Citing Clinical Trials by Country

Country	Published, 2013	Published, 2019	Change, 2013 to 2019	% Change, 2013 to 2019
China	1724	5620	3896	226.0
Poland	158	311	153	96.8
Japan	1167	2159	992	85.0
Spain	487	892	405	83.2
Italy	1241	2205	964	77.7
Austria	191	319	128	67.0
Korea	635	1045	410	64.6
Netherlands	617	1011	394	63.9
Belgium	211	335	124	58.8
France	695	1075	380	54.7
Denmark	223	343	120	53.8
Austria	597	907	310	51.9
Czech Republic	320	485	165	51.6
Brazil	286	431	145	50.7
Greece	158	228	70	44.3
Sweden	276	391	115	41.7
Germany	1555	2148	593	38.1
Canada	791	1044	253	32.0
India	302	393	91	30.1
Iran	167	215	48	28.7
Great Britain	1680	2104	424	25.2
Turkey	231	259	28	12.1
U.S.A.	11207	12536	1329	11.9

Notes: Table D.1 records the number of papers that cite clinical trials published in the preceding three years in 2013 and 2019, by country. We include all countries that publish more than 200 papers that cite clinical trials in 2019. The fourth column displays the change in the number of 'citing' papers between 2013 and 2019. The fifth column displays the percent change in the number of 'citing' papers between 2013 and 2019.

literature reviews, etc., based on the contents of their text. We reviewed all hand-labeled records from this training process and identified a set of keywords and phrases that appeared most frequently. We constructed a list of phrases and keywords that consistently indicated, to a human labeler, that a record is a meta-analysis or review article. This list is reproduced in Footnote 38.

Using this list of search terms, we search the abstracts of all records in our citing sample. We flag any record with any of the keywords listed in Footnote 38 as a meta-analysis and flag any record that mentions at least two of the databases listed in Footnote 38 as well.

The objective of this exercise is to yield a conservative estimate of the change in the quantity of meta-analyses over our period of interest. In Figure D.8, we plot this trend. Panel A plots the raw count of meta-analyses flagged by this procedure. Panel B plots the proportion of records in our citing sample that are designated as meta-analyses via this search. As expected, this approach captures a small number of records:

TABLE D.2. Composition of Publications that Cite Clinical Trials

	Published in 2013	Published in 2019
Clinical Trial	1	0
Editorial	3	2
Pre-Clinical	26	17
Non-Drug RCT	1	4
Protocol	1	1
Case Study	5	4
Observational	31	30
Meta-Analysis	32	41

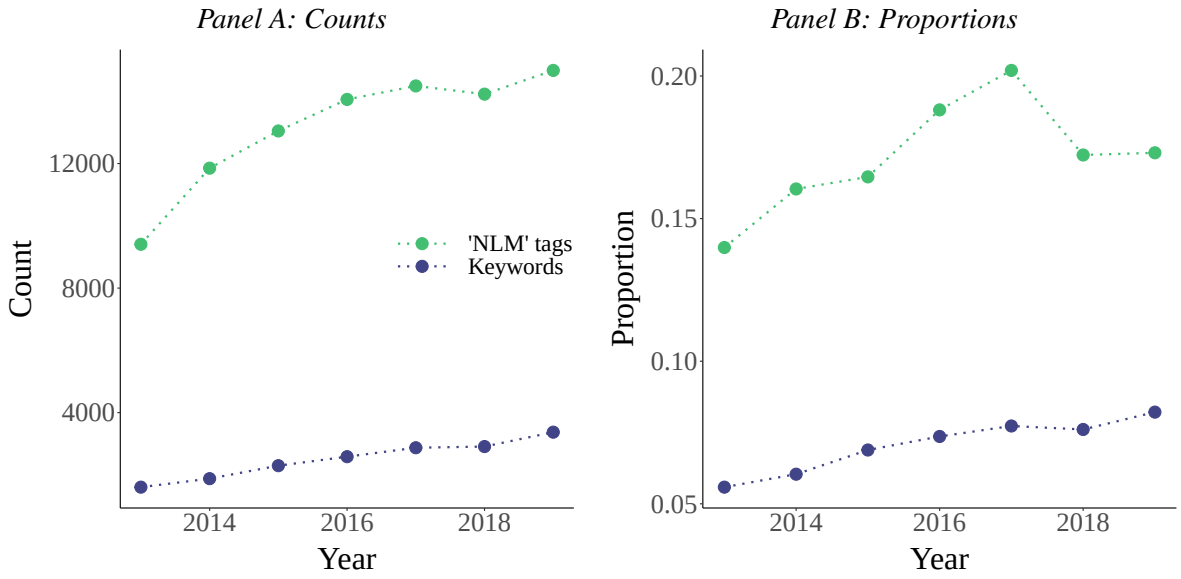
Notes:Table D.2 records the results of an exercise aimed at characterizing the composition of the set of publications that cite clinical trials (three-year sample). We randomly draw 100 publications from this sample that were published in 2013 and 2019, respectively. We categorize each publication into one of seven categories. We report the number of publications, in each year, that fall into each category.

in 2013, there are 1,606 records in our citing sample flagged as meta-analyses. This rises to 3,372 records in 2019. Although the baseline level this small, this corresponds to a 109 percent increase over this period.

D.3.3 NLM Tag Queries for Meta-Analyses. Next, we observe that NLM tags, added when publication records are indexed in PubMed, also record information that may signal whether a record is a meta-analysis. On inspection, three of the 77 NLM tags are relevant: "literature review," "meta-analysis," and "systematic review." Of course, as Section 2.3 makes clear, NLM tags may overcount records. As Appendix D.3.2 generates what is likely a lower bound, we proceed with this strategy, noting that it is likely to yield an upper bound.

We flag all records in our citing sample indexed with one of these three NLM tags. We plot the resulting trend—as a count in Panel A and as a proportion of records in our citing sample in Panel B—of Figure D.8. In 2013, there are 9,407 records satisfying these criteria. In 2019, there are 14,991 such records. Over our period of interest, there is, thus, a 59 percent increase.

FIGURE D.8. Trends in Review Articles and Meta-Analyses



Notes: Figure D.8 displays the results of two exercises aimed at characterizing the change in the quantity of meta-analyses in the set of publications that cite clinical trials (three-year sample) over time. Estimates associated with the keyword search discussed in Appendix D.3.2 are displayed in purple. Estimates associated with the 'NLM' tag query discussed in Appendix D.3.3 are displayed in purple. Panel A displays counts of the number of publications that satisfy both sets of criteria. Panel B displays proportions of publications that satisfy both sets of criteria.

APPENDIX E. PROMPT REPOSITORY

This appendix serves as a repository for figures displaying each of the prompts considered in Appendix C.3. Throughout, the dummy text "{abstract}" is placed in the location where the abstract of a publication would be input. The baseline "True/False" Prompt 1.0 is displayed in Figure E.1. Refinements to this prompt are displayed in Figures E.2 to E.4. The prompt that asks the model to categorize excluded publications, i.e., Prompt 2.0, is displayed in Figure E.5. A refinement to this prompt is given in Figure E.6. The prompt that asks the model to explain why it has excluded a publication, i.e., Prompt 3.0, is displayed in Figure E.7. In turn, a refinement to this prompt is given in Figure E.8.

FIGURE E.1. Prompt 1.0: True/False, Base

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, return `FALSE'. Do not return any extraneous text. You must return either `TRUE' or `FALSE'.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.2. Prompt 1.1: True/False, Initial Examples

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, return `FALSE'. Do not return any extraneous text. You must return either `TRUE' or `FALSE'.

For example, if the publication evaluates the effects of a surgical or medical device, procedure, or diagnostic, you should return `FALSE'. This means that if the publication studies any procedure used in the course of clinical care without also studying the effects of a specific, named drug, you should return `FALSE'. Studies of ventricular assist device placement and thrombolysis may not satisfy the criteria.

If the publication evaluates the effects of a vitamin, supplement, food/diet program, or beverage, you should return `FALSE'. This means that if the publication studies any intervention that distributes nutritional aids and supplementary sources of vitamins without also studying the effects of a specific, named drug, you should return `FALSE'. Studies of iron supplementation, omega-3 oil, infant formula, and diets that include healthy foods may not satisfy the criteria.

If the publication evaluates the effects of a behavioral therapy, physical therapy, exercise program, or other lifestyle intervention, you should return `FALSE'. This means that if the publication studies any intervention that aims to alter human behavior without also studying the effects of a specific, named drug, you should return `FALSE'. Studies of smoking cessation programs, peer counseling, parenting courses, and cardiovascular fitness programs may not satisfy the criteria.

If the publication summarizes findings of other studies only, you should return `FALSE'. This means that if the publication reports findings from a meta-analysis, systematic review, literature review, literature search, or case review without also presenting novel findings on the effects of a specific, named drug, you should return `FALSE'. Studies that reference database searches in, for example, MEDLINE or EMBASE may not satisfy the criteria.

If the publication reports results from a retrospective or observational study, you should return `FALSE'. This means that if the publication reports findings that use previously collected data, or reports results from a study in which the investigators had no control over assignment to treatment or other experimental conditions, you should return `FALSE'. Studies that reference certain designs--such as case-control studies, cohort studies, or propensity score matching studies--may not satisfy the criteria. Similarly, studies that re-analyze previously collected data, often drawn from existing databases or registries, may not satisfy the criteria.

If the publication describes a clinical trial protocol, without also reporting results from the study, you should return `FALSE'. This means that if the publication describes the design, recruitment strategy, and intervention plan for a study, but does not report any findings, you should return `FALSE'. Publications written in the future tense, which describe interventions that will happen in the future, may not satisfy the criteria. Note that a protocol publication need not explicitly use the word ``protocol.''

If the publication evaluates the effects of a drug on animals, you should return `FALSE'. This means that if the publication describes a study that enrolled any non-human participants, you should return `FALSE'. Studies of rats, swine, primates--and studies on rat, swine, and primate models or cell lines--may not satisfy the criteria.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.3. Prompt 1.2: True/False, Extended Examples, Short

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific, named investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, return `FALSE'. Do not return any extraneous text. You must return either `TRUE' or `FALSE'.

For example, if the publication summarizes or aggregates findings of other studies only, you should return `FALSE'. This means that if the publication reports findings from a meta-analysis, systematic review, literature review, literature search, case review, or post hoc evaluation of the findings of other trials without also presenting novel findings on the effects of a specific, named drug, you should return `FALSE'. Studies that reference database searches in, for example, MEDLINE or EMBASE may not satisfy the criteria.

If the publication reports results from a retrospective or observational study, you should return `FALSE'. This means that if the publication reports findings that use previously collected data, or reports results from a study in which the investigators had no control over assignment to treatment or other experimental conditions, you should return `FALSE'. Studies that reference certain designs--such as case-control studies, cohort studies, or propensity score matching studies--may not satisfy the criteria. Similarly, studies that re-analyze previously collected data, often drawn from existing databases or registries, may not satisfy the criteria.

If the publication describes a clinical trial protocol, without also reporting results from the study, you should return `FALSE'. This means that if the publication describes the design, recruitment strategy, and intervention plan for a study, but does not report any findings, you should return `FALSE'. Publications written in the future tense, which describe interventions that will happen in the future, may not satisfy the criteria. Note that a protocol publication need not explicitly use the word ``protocol.''

If the publication evaluates the effects of a drug on animals, you should return `FALSE'. This means that if the publication describes a study that enrolled any non-human participants, you should return `FALSE'. Studies of rats, swine, primates--and studies on rat, swine, and primate models or cell lines--may not satisfy the criteria.

If the publication evaluates the effects of a surgical or medical device, procedure, or diagnostic, you should return `FALSE'. This means that if the publication studies any procedure used in the course of clinical care without also studying the effects of a specific, named drug, you should return `FALSE'. Studies of ventricular assist device placement and thrombolysis may not satisfy the criteria.

If the publication evaluates the effects of a vitamin, supplement, food/diet program, or beverage, you should return `FALSE'. This means that if the publication studies any intervention that distributes nutritional aids and supplementary sources of vitamins without also studying the effects of a specific, named drug, you should return `FALSE'. Studies of iron supplementation, omega-3 oil, infant formula, probiotic supplements, and diets that include healthy foods may not satisfy the criteria. Note that these supplements may have different names in different settings (e.g., "ferrous" supplements or "iron" complexes).

If a publication only mentions a drug or medicine, but does not evaluate its effects, you should return `FALSE'. This means that if the publication studies an intervention in patients who happen to be taking a drug or measures the effects of a naturally-occuring substance in the human body, you should return `FALSE’.

If the publication evaluates the effects of a behavioral therapy, physical therapy, exercise program, or other lifestyle intervention, you should return `FALSE'. This means that if the publication studies any intervention that aims to alter human behavior without also studying the effects of a specific, named drug, you should return `FALSE'. Studies of smoking cessation programs, peer counseling, parenting courses, and cardiovascular fitness programs may not satisfy the criteria.

FIGURE E.3. Prompt 1.2: True/False, Extended Examples, Short (Cont.)

To determine if a drug is ``specific'' and ``named,'' you should consider whether it is referenced by a unique brand, chemical, generic, or internal company name. General classes of drugs or therapies do not meet this requirement. For example, studies of "statins", "antidepressants", or "anesthesia" refer to generic categories of drugs, not specific drugs and would not meet this requirement. Similarly, "antihypertensive therapies", "antiretrovirals", and "dopaminergic therapies" are generic classes and would not meet this requirement. For this requirement to be met, you should be able to name at least one drug or medicine that is being evaluated in this study.

Please consider these guidelines very carefully before returning an answer. Before returning an answer, please identify the specific name of at least one drug or medicine being studied. Confirm that this named medicine/drug is not a food, vitamin, or supplement. Confirm, also, that this is not a "class" or "type" of medicine without a specific name.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.4. Prompt 1.3: True/False, Extended Examples, Long

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific, named investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, return `FALSE'. Do not return any extraneous text. You must return either `TRUE' or `FALSE'.

For example, if the publication summarizes, synthesizes, or aggregates findings of other studies only, you should return `FALSE'. This means that if the publication reports findings from a meta-analysis, systematic review, literature review, literature search, case review, or post hoc evaluation of the findings of other trials without also presenting novel findings on the effects of a specific, named drug, you should return `FALSE'. For example, studies that reference database searches in, for example, MEDLINE or EMBASE may not satisfy the criteria. Articles intended to summarize the state of treatment guidelines or provide guidelines for clinicians may not satisfy the criteria.

If the publication reports results from a retrospective or observational study, you should return `FALSE'. This means that if the publication reports findings that use previously collected data, reports a post-hoc analysis, or reports results from a study in which the investigators had no control over assignment to treatment or other experimental conditions, you should return `FALSE'. For example, studies that reference certain designs--such as case-control studies, cohort studies, or propensity score matching studies--may not satisfy the criteria. Similarly, studies that re-analyze previously collected data, often drawn from existing databases or registries, may not satisfy the criteria. Finally, studies that explicitly describe their design as "retrospective" or "observational" do not satisfy the criteria.

If the publication describes a clinical trial protocol, without also reporting results from the study, you should return `FALSE'. This means that if the publication describes the design, recruitment strategy, and intervention plan for a study, but does not report any findings, you should return `FALSE'. For example, publications written in the future tense, which describe interventions that will happen in the future, may be protocols. Similarly, articles that describe results that "will be" collected are likely to be protocols. Note that a protocol publication need not explicitly use the word ``protocol.''

If the publication evaluates the effects of a drug on any animal subjects, you should return `FALSE'. This means that if the publication describes a study that enrolled ANY non-human participants, you should return `FALSE'. For example, studies of rats, swine, primates--and studies on rat, swine, and primate models or cell lines--may not satisfy the criteria. Any other study on animals, similarly, will not satisfy the criteria.

If the publication evaluates the effects of a surgical or medical device, procedure, or diagnostic, you should return `FALSE'. This means that if the publication studies any procedure used in the course of clinical care without also studying the effects of a specific, named drug, you should return `FALSE'. For example, studies of ventricular assist device placement and thrombolysis may not satisfy the criteria.

If the publication evaluates the effects of a vitamin, supplement, food/diet program, or beverage, you should return `FALSE'. This means that if the publication studies any intervention that distributes nutritional aids and supplementary sources of vitamins without also studying the effects of a specific, named drug, you should return `FALSE'. For example, studies of iron supplementation, omega-3 oil, infant formula, probiotic supplements, and diets that include healthy foods may not satisfy the criteria. Note that these supplements may have different names in different settings (e.g., "ferrous" supplements or "iron" complexes).

If the publication only reports the results of a scientific study conducted in a laboratory setting (i.e., not on live human subjects), you should return `FALSE'. This means that if the publication studies a setting with no human subjects--including a laboratory investigation of cell lines or biological materials drawn from human subjects--you must return `FALSE'.

FIGURE E.4. Prompt 1.3: True/False, Extended Examples, Long (Cont.)

If a publication only mentions a drug or medicine, but does not evaluate its effects, you should return `FALSE'. This means that if the publication studies an intervention in patients who happen to be taking a drug or measures the effects of a naturally-occuring substance in the human body, you should return `FALSE'.

If the publication evaluates the effects of a behavioral therapy, physical therapy, exercise program, or other lifestyle intervention, you should return `FALSE'. This means that if the publication studies any intervention that aims to alter human behavior without also studying the effects of a specific, named drug, you should return `FALSE'. For example, studies of smoking cessation programs, peer counseling, parenting courses, and cardiovascular fitness programs may not satisfy the criteria.

To determine if a drug is ``specific'' and ``named,'' you should consider whether it is referenced by a unique brand, chemical, generic, or internal company name. General classes of drugs or therapies do not meet this requirement. For example, studies of "statins", "antidepressants", or "anesthesia" refer to generic categories of drugs, not specific drugs and would not meet this requirement. Similarly, "antihypertensive therapies", "antiretrovirals", and "dopaminergic therapies" are generic classes and would not meet this requirement. For this requirement to be met, you should be able to name at least one drug or medicine that is being evaluated in this study. Note, however, that simply because a medicine name is mentioned does not mean that these criteria are satisfied. The remainder of the definition must still be satisfied.

Please consider these guidelines very carefully before returning an answer. Before returning an answer, please identify the specific name of at least one drug or medicine being studied. Confirm that this named medicine/drug is not a food, vitamin, or supplement. Confirm, also, that this is not a "class" or "type" of medicine without a specific name.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.5. Prompt 2.0: Categorize Exclusion Restriction

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, you must provide at least one reason for your choice. To do so, you must follow the following directions.

Reasons:
- Return "NO DRUG" if the publication does not report a clinical trial that studies at least one medicine/drug dispensed in the course of medical care. Drugs MAY include the following: investigational compounds, biological therapies (including vaccines), and small-molecule, synthetic products. Drugs DO NOT include the following: dietary supplements, diets, behavioral interventions, medical and surgical procedures, medical and surgical devices, diagnostic tests, and dental substances. Drugs also do not include chemicals, substances, and toxins that are not dispensed in the course of medical care.
- Return "ANIMAL" if the publication reports the results of a study conducted, in part or wholly, on animals.
- Return "NO HUMAN SUBJECTS" if the publication only reports the results of a scientific study conducted in a laboratory setting (i.e., not on live human subjects). Studies with no human subjects include laboratory investigations of cell lines or biological materials drawn from human subjects if no living human subjects are actively involved.
- Return "META-ANALYSIS" if the publication only reports the results of a meta-analysis, literature review, literature search, narrative review, or summary of individuals' careers. Summaries of previously published clinical trials--without new data collection or analysis--should be excluded for this reason.
- Return "OBSERVATIONAL" if the publication only reports results from an observational study (i.e., there was no active intervention). Observational studies may include studies using data drawn from existing datasets, registries, and electronic health records.
- Return "RETROSPECTIVE" if the publication only reports results from a retrospective study. A retrospective study is a non-interventional study that compares two groups of people: those with the disease or condition under study (cases) and a very similar group of people who do not have the disease or condition (controls). The allocation of people to the cases or controls is not chosen by the researcher, and data may be drawn from a dataset not explicitly intended for the research being reported.
- Return "PROTOCOL" if the publication only reports the protocol for a clinical trial. A protocol is a formal analysis plan for a clinical trial that describes the design and attributes of the study, but does not included clinical results (e.g., no statistical results or analyses are reported).
- Return "CASE REPORT" if the publication only reports information collected from clinical cases. A case report is a detailed report of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient or set of patients.
- Return "NOT PUBLISHED IN ENGLISH" if the publication is published in a language other than English.

If the publication does not satisfy the criteria for multiple reasons, return a complete list of reasons separated by semicolons, e.g., "OBSERVATIONAL; NO DRUG". If the publication satisfies all classification criteria, only return `TRUE'. Do not return any extraneous text. You must return either `TRUE' or a complete list of reasons described above.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.6. Prompt 2.1: Categorize Exclusion Restriction, Examples

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE'. If the publication does not satisfy all criteria, you must provide at least one reason for your choice. To do so, you must follow the following directions.

Reasons:
- Return "NO DRUG" if the publication does not report a clinical trial that studies at least one medicine/drug dispensed in the course of medical care. Drugs MAY include the following: investigational compounds, biological therapies (including vaccines), and small-molecule, synthetic products. Drugs DO NOT include the following: dietary supplements, diets, behavioral interventions, medical and surgical procedures, medical and surgical devices, diagnostic tests, and dental substances. Drugs also do not include chemicals, substances, and toxins that are not dispensed in the course of medical care.
- Return "ANIMAL" if the publication reports the results of a study conducted, in part or wholly, on animals.
- Return "NO HUMAN SUBJECTS" if the publication only reports the results of a scientific study conducted in a laboratory setting (i.e., not on live human subjects). Studies with no human subjects include laboratory investigations of cell lines or biological materials drawn from human subjects if no living human subjects are actively involved.
- Return "META-ANALYSIS" if the publication only reports the results of a meta-analysis, literature review, literature search, narrative review, or summary of individuals' careers. Summaries of previously published clinical trials--without new data collection or analysis--should be excluded for this reason.
- Return "OBSERVATIONAL" if the publication only reports results from an observational study (i.e., there was no active intervention). Observational studies may include studies using data drawn from existing datasets, registries, and electronic health records.
- Return "RETROSPECTIVE" if the publication only reports results from a retrospective study. A retrospective study is a non-interventional study that compares two groups of people: those with the disease or condition under study (cases) and a very similar group of people who do not have the disease or condition (controls). The allocation of people to the cases or controls is not chosen by the researcher, and data may be drawn from a dataset not explicitly intended for the research being reported.
- Return "PROTOCOL" if the publication only reports the protocol for a clinical trial. A protocol is a formal analysis plan for a clinical trial that describes the design and attributes of the study, but does not included clinical results (e.g., no statistical results or analyses are reported).
- Return "CASE REPORT" if the publication only reports information collected from clinical cases. A case report is a detailed report of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient or set of patients.
- Return "NOT PUBLISHED IN ENGLISH" if the publication is published in a language other than English.

If the publication does not satisfy the criteria for multiple reasons, return a complete list of reasons separated by semicolons, e.g., "OBSERVATIONAL; NO DRUG". If the publication satisfies all classification criteria, only return `TRUE'. Do not return any extraneous text. You must return either `TRUE' or a complete list of reasons described above.

For example, if the publication evaluates the effects of a surgical or medical device, procedure, or diagnostic, you should return `NO DRUG'. This includes publications that study any procedure used in the course of clinical care without also studying the effects of a specific, named drug. This may include, for example, studies of ventricular assist device placement and thrombolysis.

FIGURE E.6. Prompt 2.1: Categorize Exclusion Restriction, Examples (Cont.)

If the publication evaluates the effects of a vitamin, supplement, food/diet program, or beverage, you should also return `NO DRUG'. This includes publications that study any intervention that distributes nutritional aids and supplementary sources of vitamins without also studying the effects of a specific, named drug. This may include, for example, studies of iron supplementation, omega-3 oil, infant formula, and diets that include healthy foods.

If the publication evaluates the effects of a behavioral therapy, physical therapy, exercise program, or other lifestyle intervention, you should also return `NO DRUG'. This includes publications that study any intervention that aims to alter human behavior without also studying the effects of a specific, named drug. This may include, for example, studies of smoking cessation programs, peer counseling, parenting courses, and cardiovascular fitness programs.

If the publication summarizes findings of other studies only, you should return `META-ANALYSIS'. This includes publications that report findings from a meta-analysis, systematic review, literature review, literature search, or case review without also presenting novel findings on the effects of a specific, named drug. This may include, for example, studies that reference database searches in, for example, MEDLINE or EMBASE.

If the publication reports results from a retrospective study, you should return `RETROSPECTIVE'. This includes publications that report findings that use previously collected data. This may include, for example, studies that re-analyze previously collected data, often drawn from existing databases or registries.

If the publication reports results from an observational study, you should return `OBSERVATIONAL'. This includes publications that report findings that use previously collected data, or report results from a study in which the investigators had no control over assignment to treatment or other experimental conditions. This may include, for example, studies that reference certain designs--such as case-control studies, cohort studies, or propensity score matching studies.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.7. Prompt 3.0: Provide Reason for Exclusion

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE.' If the publication does not satisfy all criteria, you must provide a reason for your choice. In particular, you must return `FALSE' followed by an explanation for why it does not meet the criteria. Your response should take the form: `FALSE: [xxx].' Do not return any extraneous text. You must return either `TRUE' or `FALSE: ', followed by your explanation.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer:

FIGURE E.8. Prompt 3.1: Provide Reason for Exclusion, Examples

At the end of this prompt, you will be shown an abstract from an academic publication indexed in the PubMed/ MEDLINE database.

Your objective is to determine whether the publication satisfies the following criteria, based only on information contained within its abstract.

Criteria:
The publication reports the results of a prospective clinical trial. The clinical trial may be of any phase. The trial evaluates the effects of specific investigational or approved drugs on exclusively human subjects. The abstract is written in English.

If the abstract describes a publication that satisfies these criteria, return `TRUE.' If the publication does not satisfy all criteria, you must provide a reason for your choice. In particular, you must return `FALSE' followed by an explanation for why it does not meet the criteria. Your response should take the form: `FALSE: [xxx].' Do not return any extraneous text. You must return either `TRUE' or `FALSE: ', followed by your explanation.

For example, if the publication evaluates the effects of a surgical or medical device, procedure, or diagnostic, you should return `FALSE: ' followed by an explanation that identifies the device, procedure, or diagnostic being studied and specifies that it is not a drug. This includes publications that study any procedure used in the course of clinical care without also studying the effects of a specific, named drug. This may include, for example, studies of ventricular assist device placement and thrombolysis.

If the publication evaluates the effects of a vitamin, supplement, food/diet program, or beverage, you should return `FALSE: ' followed by an explanation that identifies the vitamin, supplement, food/diet program, or beverage being studied and specifies that it is not a drug. This includes publications that study any intervention that distributes nutritional aids and supplementary sources of vitamins without also studying the effects of a specific, named drug. This may include, for example, studies of iron supplementation, omega-3 oil, infant formula, and diets that include healthy foods.

If the publication evaluates the effects of a behavioral therapy, physical therapy, exercise program, or other lifestyle intervention, you should return `FALSE: ' followed by an explanation that identifies the behavioral therapy, physical therapy, exercise program, or lifestyle intervention being studied and specifies that it is not a drug. This includes publications that study any intervention that aims to alter human behavior without also studying the effects of a specific, named drug. This may include, for example, studies of smoking cessation programs, peer counseling, parenting courses, and cardiovascular fitness programs.

If the publication summarizes findings of other studies only, you should return `FALSE: ' followed by an explanation with two components. First, specify that the publication does not report the results of a specific prospective clinical trial and, second, provide a brief summary of what the publication does report. This includes publications that report findings from a meta-analysis, systematic review, literature review, literature search, or case review without also presenting novel findings on the effects of a specific, named drug. This may include, for example, studies that reference database searches in, for example, MEDLINE or EMBASE.

If the publication reports results from a retrospective or observational study, you should return `FALSE: ' followed by an explanation with two components. First, specify that the publication does not report the results of a prospective, interventional clinical trial and, second, provide a brief summary of what the publication does report. This includes publications that report findings that use previously collected data, or report results from a study in which the investigators had no control over assignment to treatment or other experimental conditions. This may include, for example, studies that reference certain designs--such as case-control studies, cohort studies, or propensity score matching studies. Similarly, it may include studies that re-analyze previously collected data, often drawn from existing databases or registries.

If the publication describes a clinical trial protocol, without also reporting results from the study, you should return `FALSE: ' followed by an explanation that the publication does not report the results of a clinical trial. This includes publications that describe the design, recruitment strategy, and intervention plan for a study, but do not report any findings. This may include publications written in the future tense, which describe interventions that will happen in the future. Note that a protocol publication need not explicitly use the word ``protocol.''

FIGURE E.8. Prompt 3.1: Provide Reason for Exclusion, Examples (Cont.)

If the publication evaluates the effects of a drug on animals, you should return `FALSE: ' followed by an explanation that describes that the study was conducted on animals and identifies the animals being studied. This includes publications that describe studies that enrolled any non-human participants. This may include studies of rats, swine, primates--and studies on rat, swine, and primate models or cell lines.

The abstract that you will consider is as follows:

Abstract: {abstract}

Answer: