Journal of Development Economics 184 (2027) 103892

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular Article

Exploitation vs. Exploration: How knowledge spillovers shape healthcare practices and innovation☆

Yinan Liu [a,d], Chen Lin [a,*], Zixiao Li [b,c], Jing Zhang [c]

[a] School of Applied Economics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, China
[b] Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, No. 119 South, Fourth Ring West Road, Fengtai District, Beijing 100070, China
[c] China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, No. 119 South Fourth Ring West Road, Fengtai District, Beijing 100070, China
[d] Center for Strategic Studies on Fundamental Research and Chinese Modernization, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, China

ARTICLE INFO

JEL classification:
H00
I18
O38
D24

Keywords:
Knowledge spillovers
Physician behavior
Innovation in healthcare
Large language model

ABSTRACT

Improving public service quality is a central challenge in developing countries. A common policy tool is to transfer expertise and protocols to lower-tier institutions, with the expectation that such spillovers improve both current performance and long-term innovative capacity. We study this policy logic using a Chinese reform that mandated knowledge transfer from elite to lower-tier hospitals. Our analysis combines a novel measure of lower-tier hospitals' alignment with elite hospitals' clinical practice, developed with a large language model applied to clinical texts and prescription decisions, with measures of physicians' exploratory activity based on academic publications. Using a difference-in-differences design, we find that the policy increased practice alignment at targeted hospitals by 0.29 standard deviations, yielding immediate efficiency gains. At the same time, this gain in exploitation was accompanied by a decline in exploration: physician academic output fell, with publication counts declining by 7% and citation impact by 13%. Our findings highlight that policies promoting existing knowledge can close practice gaps and improve care efficiency, while also creating a potential trade-off with professionals' incentives to generate new knowledge.

1. Introduction

Persistent productivity dispersion between leading and lagging service providers is a salient feature of many developing economies, often motivating government intervention (Hsieh and Klenow, 2009; Banerjee et al., 2020; Han and Wu, 2024). A common policy response is to promote the diffusion of "best practices" from national leaders to local laggards. This approach is premised on the idea that facilitated knowledge spillovers can bridge persistent capability gaps, enhance efficiency, and improve the quality of essential services (Bloom and Van Reenen, 2007; Acemoglu et al., 2012; Bloom et al., 2013; Cai and Szeidl, 2018). This strategy is particularly prevalent in crucial public sectors like healthcare, where learning from elite centers is seen as a direct path to better patient outcomes and lower costs in resource-constrained settings (Chandra and Staiger, 2007; Chandra and Skinner, 2012; Finkelstein et al., 2016; Currie and MacLeod, 2017). Yet, this intense focus on exploiting existing knowledge may harbor a hidden trade-off. By encouraging convergence toward established routines, do such policies inadvertently suppress the local experimentation that fuels long-term, endogenous progress?

This paper investigates this fundamental tension between exploiting existing knowledge (exploitation) and fostering autonomous innovation (exploration). The trade-off, first formalized by March (1991), is central to understanding how sustainable development is achieved. Exploitation—the refinement and application of current knowledge—offers immediate gains in efficiency and reliability. Exploration—the

☆ We are deeply grateful to the Editor and two anonymous referees for their constructive comments and suggestions. We also thank Hanming Fang, Jin Feng, Binzhen Wu, and Junjian Yi, as well as seminar participants at Peking University, Tsinghua University, Fudan University, Renmin University of China, Nanjing University, Shanghai University of Finance and Economics, and Jinan University for their invaluable feedback. Excellent research assistance was provided by Mengjie Deng, Peiyao Lv, and Zhicheng Zhang. Chen Lin acknowledges financial support from the Major Program of the National Social Science Fund of China (Grant No. 22&ZD092). Zixiao Li acknowledges financial support from the Noncommunicable Chronic Diseases—National Science and Technology Major Project (Grant Nos. 2025ZD0546500 and 2025ZD0546504). All remaining errors are our own.

* Corresponding author.
E-mail addresses: yinanliu@ruc.edu.cn (Y. Liu), c_lin@ruc.edu.cn (C. Lin), lizixiao@ncrcnd.org.cn (Z. Li), zhangjing82@bjtth.org (J. Zhang).

https://doi.org/10.1016/j.jdeveco.2026.103892
Received 2 September 2025; Received in revised form 21 July 2026; Accepted 21 July 2026
Available online 31 July 2026
0304-3878/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

search for novel, potentially superior, alternatives—is uncertain but essential for building resilience and long-term, self-sustaining capabilities. This trade-off manifests with particular urgency in the healthcare sectors of developing countries. Amid resource constraints, local physicians face strong incentives to apply externally-developed standards (exploit), while the need to discover cost-effective ways to treat populations with specific local health burdens places a high premium on exploration. A policy that heavily incentivizes the former may induce dependence on established practices, making physicians less likely to deviate from externally validated protocols even when experimentation could generate valuable local knowledge (Molitor, 2018; Cutler et al., 2019; Jin et al., 2024). This raises a critical question for economic policy: can we promote the benefits of best-practice adoption without sacrificing the engine of future innovation?

To provide causal evidence on this question, we leverage China's Urban Medical Alliance (UMA) policy as a natural experiment. Launched by the central government in June 2019, the UMA policy was implemented simultaneously in a set of designated pilot areas. It mandated formal partnerships between tertiary (higher-tier) hospitals and secondary (lower-tier) hospitals, with the explicit goal of transferring clinical expertise and improving care quality. This design creates a clean quasi-experimental setting. We employ a canonical difference-in-differences (DID) framework, comparing outcomes in secondary hospitals located in UMA pilot areas (the treated group) with those in non-pilot areas (the control group) before and after the policy's introduction. Our identification strategy hinges on the fact that the policy significantly intensifying knowledge spillovers from tertiary to secondary hospitals in treated areas, while assuming that, absent the policy, trends in clinical practices and innovation would have evolved similarly across treated and non-treated areas.

Measuring the two sides of this trade-off presents significant empirical challenges. First, to capture exploitation, we need a granular measure of how closely a physician's practice style aligns with a "best-practice" benchmark. Existing studies often rely on coarse administrative codes, which miss the rich clinical nuance of medical decisions. We overcome this by developing a novel methodology based on a large language model (LLM). We collect a large dataset of radiology reports and associated prescriptions from secondary hospitals and a top-tier benchmark tertiary hospital. Using SimCSE (Simple Contrastive Learning of Sentence Embeddings), a contrastive-learning method for constructing sentence embeddings, we fine-tune our language model to transform each radiology report into a rich, contextual embedding. This allows us to match each patient case from a secondary hospital to its most clinically similar counterpart in the benchmark hospital based on the textual content of their radiology reports. We then compute the similarity of the prescriptions for these matched pairs. This yields a continuous, case-adjusted measure of practice style alignment.

Second, to capture exploration, we use physicians' academic publications to measure their engagement in formal research and new knowledge production. While other measures like patents or technology adoption are valuable, publications are particularly well-suited to capture the conceptual and clinical advances made by practicing physicians, which may not be patentable but are fundamental to scientific progress (Adams, 1990; Murray and Stern, 2007; Aghion et al., 2008; Williams, 2013). Crucially, we move beyond a simple count of articles to assess their quality. Following standard practice in the economics of innovation, we measure quality by tracking publications in top-tier journals and their subsequent citation impact (Jaffe et al., 1993; Azoulay et al., 2011; Jaravel et al., 2018). These distinct measures allow us to separately estimate the policy's effects on both the quantity and the quality of physicians' research output.

Our DID estimates reveal a trade-off. On one hand, the UMA policy succeeded in promoting exploitation: we find a significant increase in the alignment of prescribing practices between treated secondary hospitals and the benchmark hospital. This convergence was accompanied by a significant decrease in medical spending, suggesting that the adoption of best practices also enhanced efficiency. On the other hand, these gains were accompanied by a decline in exploration. We document a decline in the academic output of physicians in treated secondary hospitals. This reduction was not merely in quantity; we find a significant drop in the number of publications in top-tier journals and in the citation impact of their work. Taken together, these results suggest that promoting the adoption of established clinical knowledge can improve current practice while shifting physicians away from the production of new academic knowledge.

This study contributes to the literature in three ways. First, this study provides causal evidence of the exploitation–exploration trade-off at the individual professional level. While this trade-off is a central tenet of innovation economics (March, 1991), direct causal evidence for individuals is scarce. The literature has predominantly focused on firm-level strategic decisions, leaving the individual-level response largely unexplored (Levinthal and March, 1993; Rothaermel and Deeds, 2004; Aghion et al., 2005; Akcigit and Kerr, 2018). We isolate a policy designed to diffuse a "best practice" and show that it increases reliance on established routines while reducing research activity among recipient professionals. Recent research has begun to apply the exploration framework to individual professionals, with a notable focus on selection challenges like hiring (Li et al. forthcoming). Our study contributes to this emerging literature by shifting the analytical focus from talent selection to the subsequent challenge of creative production: we investigate how professionals allocate effort between adopting established practices and producing new knowledge once a "best practice" is readily available. Ultimately, our findings speak to a broader tension in economic policy: interventions that improve short-run performance may also weaken incentives for experimentation and new knowledge creation (Holmström, 1999; Manso, 2011; Acemoglu and Akcigit, 2012; Aghion et al., 2015).

Second, we contribute to the broader economic literature on path dependency by showing how externally supplied standards can reshape the balance between exploitation and exploration within recipient organizations. Economic theories have provided powerful explanations for "lock-in", largely focusing on two types of self-reinforcing dynamics: increasing returns at the market level that entrench a dominant technology (Arthur, 1989; Rochet and Tirole, 2003; Clements and Ohashi, 2005; Halaburda et al., 2020), and inertia and learning at the individual level that create state dependence on one's own past actions (Handel, 2013; Jin et al., 2024). Our study shifts the focus from these self-referential processes to a policy-induced form of dependence on externally validated practices. Specifically, we show that a policy designed to institutionalize a "best practice" induced convergence toward the benchmark and was accompanied by a decline in research output, particularly in domains most closely connected to the transferred knowledge. These findings are consistent with a localized form of protocol dependence, in which externally validated standards may narrow the scope or incentives for independent inquiry. More broadly, our findings speak to emerging concerns about cognitive dependence in AI-assisted decision-making, whereby reliance on externally generated guidance may improve immediate performance while narrowing the scope for independent experimentation and knowledge creation (Brynjolfsson et al., 2025; Kosmyna et al., 2025).

Finally, our study contributes a novel approach to measuring practice style alignment using text as data. Quantifying how practitioners make decisions is a central challenge for a vast range of economic questions, from the study of knowledge spillovers to organizational learning. Yet, practice styles are complex and difficult to observe directly, often forcing researchers to rely on coarse administrative records as proxies (Abaluck et al., 2016; Finkelstein et al., 2016; Currie and MacLeod, 2017; Molitor, 2018). Inspired by recent progress in text analysis in economics (Gentzkow et al., 2019; Chen et al., 2024; Fang et al., 2025; Juhász et al., 2025), we contribute to this frontier by developing a scalable method based on recent advances in natural language processing. By fine-tuning a pre-trained language-model encoder on unstructured

medical reports, we generate a high-dimensional, case-adjusted measure of practice alignment. This text-as-data methodology not only enables our main analysis but also offers a scalable and replicable approach for future research on the diffusion of complex professional knowledge, peer effects, and organizational culture in healthcare and other knowledge-intensive settings.

The remainder of this paper is organized as follows: Section 2 describes disparities in China's healthcare system, physicians' publication incentives, and details of the UMA policy. Section 3 introduces our primary data sources and details the novel construction of our key outcome variable for exploitation, namely the practice similarity score, using language-model embeddings. Section 4 presents the empirical strategy. Section 5 presents the main effects on practice alignment, patient outcomes, and academic output. Section 6 examines heterogeneity in the response, and Section 7 provides robustness and placebo checks. Section 8 discusses the implications of the findings and offers concluding remarks.

2. Institutional background

2.1. Physician expertise disparities in China's three-tier healthcare system

China's healthcare system operates under a three-tier structure, formally established in 1989 to improve the organization and delivery of medical services (Ministry of Health of the People's Republic of China, 1989). At the top of this hierarchy are tertiary hospitals, which play a crucial role in managing the most complex and critical health conditions. These hospitals, typically concentrated in urban centers, benefit from superior funding, advanced medical technologies, and institutional prestige (Lu et al., 2019; Hu et al., 2021). These advantages allow tertiary hospitals to offer highly competitive salaries and better working conditions, enabling them to recruit the most skilled and specialized physicians, including those with advanced training or international experience (Yip and Hsiao, 2008; Pan et al., 2016). As a result, tertiary hospitals attract top-tier medical talent, which consolidates their position as leaders in both clinical practice and medical innovation.

In contrast, non-tertiary hospitals, which encompass both secondary hospitals and primary medical institutions, face substantial challenges in attracting and retaining similarly qualified physicians (Li et al., 2017, 2020). While secondary hospitals serve as regional healthcare hubs providing essential services to broader communities, and primary institutions focus on delivering basic healthcare, both tiers are constrained by limited financial resources, less access to advanced medical technology, and reduced institutional prestige compared to tertiary hospitals. These constraints make it difficult for non-tertiary hospitals to compete for top medical talent, resulting in a workforce with less specialized training and experience. As illustrated in Figure A1, non-tertiary institutions, such as community health centers, employ a disproportionately lower share of physicians with graduate or bachelor-level degrees, while a significant portion of their workforce holds vocational or high school qualifications. Consequently, the physicians at non-tertiary hospitals, while providing necessary care, generally have lower technical proficiency compared to their peers in tertiary hospitals, further exacerbating the gap in healthcare quality between these tiers. This disparity in expertise creates the very conditions that policies promoting best-practice diffusion are designed to address, setting the stage for a potential increase in knowledge exploitation.

2.2. The growing role of academic publishing in medical research in China

The disparities in talent distribution not only influence the quality of clinical care but also shape the landscape of medical exploration, for which academic publishing serves as a primary, albeit imperfect, measure. In the medical field, publishing is a critical mechanism for disseminating novel findings, validating new techniques, and documenting unusual clinical cases, thereby contributing to the collective knowledge base (Vandenbroucke, 1998; Furman and Stern, 2011; Williams, 2013). The production of this new knowledge, however, inherently reflects the disparities of the three-tier system. Positioned at the top of the healthcare hierarchy, tertiary hospitals unequivocally dominate national academic output and generate the vast majority of medical research (Li et al., 2022). Conversely, secondary hospitals contribute substantially less to high-level academic publishing (Cheng et al., 2021; Hao et al., 2023).

However, a less obvious pattern challenges the simple narrative of tertiary-hospital dominance in research: although secondary hospitals lag behind tertiary hospitals in high-level outputs, they still generate a sizable volume of routine medical publications (Hao et al., 2023). This counterintuitive productivity is not accidental, but reflects systemic policy shifts that made research performance relevant to both institutional evaluation and individual career advancement. Since the early 2000s, publication requirements have been increasingly incorporated into physician promotion criteria, while scientific research indicators have become more salient in hospital evaluations and ranking systems, effectively institutionalizing publishing as a measure of professional competence (Yuan et al., 2013; Zhang et al., 2014; Dong et al., 2021; Li et al., 2022).[1]

These publication-linked career incentives encouraged "exploration" among physicians in secondary hospitals, leading to a measurable baseline of innovation even within these resource-constrained institutions.[2] This intense publishing pressure only started to ease in 2021, when the Health Professional Title Reform sought to reduce the rigid reliance on publications in physician promotion. However, the reform did not remove academic output from evaluation altogether: publications could still be submitted as representative works, and local implementation was gradual.[3] Thus, during our main sample period, secondary-hospital physicians continued to face meaningful publication incentives. Against this backdrop, the 2019 alliance-based reform analyzed below therefore allows us to test whether promoting best-practice diffusion inadvertently crowds out this pre-existing margin of innovation by acting as a strong shock to the returns on exploitation.[4]

2.3. The urban medical alliance policy

Beyond shaping academic outputs, this structural imbalance in physician expertise and resource allocation between tertiary and non-tertiary hospitals has significant implications for clinical patient load (Pan et al., 2015; Lu et al., 2019). Although the tiered system is

[1] A common justification for involving physicians in research is that research engagement can help clinicians evaluate clinical practice more systematically and contribute to evidence-based care (Jacob, 2016). In China, however, this rationale coexisted with strong career incentives: publication requirements in promotion and hospital-level evaluation generated substantial publication pressure among clinicians.

[2] Due to resource constraints, physicians in secondary hospitals often focus on smaller-scale research, such as case reports or localized studies, which are more feasible than large-scale clinical trials. See Tsauo (2013) for evidence of publication pressure affecting career paths, and Li et al. (2022) for the role of hospital rankings in driving research output.

[3] The 2021 national reform emphasized breaking the practice of evaluating physicians solely by publications and allowed clinical case records and other representative works to count in title review. Local implementation was staged; for example, Beijing's local implementation rules were issued in 2023 and listed papers, case records, technical reports, clinical trial reports, and other materials as possible representative works. See https://www.gov.cn/zhengce/zhengceku/2021-08/05/content_5629566.htm and https://news.cyol.com/gb/articles/2023-03/31/content_qbePJ4CpdJ.html.

[4] We return to the 2021 title reform in Section 7 and show that the main publication results are robust to controlling for this concurrent change in career incentives.

designed to allocate responsibilities based on the complexity of medical cases, in practice, patients are not required to adhere to the hierarchy when seeking care. Many patients, even those with less severe conditions, bypass lower-tier institutions and go directly to tertiary hospitals, drawn by their superior reputation and resources (Lu et al., 2019). This tendency has resulted in overcrowding in tertiary hospitals. Outpatient visits handled by tertiary hospitals have nearly doubled from 28.6% in 2005 to 54.1% in 2020, despite the fact that they only account for 8.5% of all hospital institutions. This concentration of care is also evident in inpatient services: in 2019, bed occupancy rates were 97.5% for tertiary hospitals, compared to 81.6% for secondary hospitals and 54.7% for primary hospitals.[5]

To address the severe inefficiencies driven by these disparities, the Chinese government has rolled out a series of structural healthcare reforms. Chief among these is the Urban Medical Alliance (UMA) policy, launched nationally in June 2019 to structurally integrate regional healthcare delivery.[6] Unlike previous reforms that prioritized patient triage or strict price controls, the UMA operates fundamentally through knowledge transfer across healthcare tiers.[7] By establishing alliances between tertiary and lower-tier hospitals, the policy seeks to elevate the clinical capabilities of secondary institutions through structured clinical mentorship and training. The policy operates primarily through the downward transfer of clinical expertise—what Chinese policy documents describe as the lead hospital's *jishu fushe* (expertise diffusion) role—rather than merely redirecting patients across tiers. In practice, this support takes the form of physician dispatch, hands-on clinical training, joint case consultations, mentorship, telemedicine for real-time consultation, and joint development of clinical specialty services. These channels repeatedly expose lower-tier physicians to tertiary hospitals' diagnostic routines, treatment protocols, and clinical management standards, effectively acting as an exogenous shock that lowers the cost of adopting established best practices and shifts incentives toward knowledge exploitation.[8]

The UMA policy was implemented in 118 pilot cities across 31 provinces. Each province selected several cities or districts based on criteria set by the National Health Commission, which emphasized strong municipal government support for healthcare reforms, well-established healthcare infrastructure, and demonstrated capacity for collaboration between institutions.[9] Within each city, a leading tertiary hospital anchors the alliance, with local secondary and primary hospitals participating as member institutions.

Official statistics capture the scale of this downward transfer of clinical expertise. By the end of 2019, tertiary hospitals had supported the implementation of 15,656 new clinical technologies and projects in lower-level hospitals, representing a 34.5% increase from 2018, and dispatched 780,000 professionals to secondary hospitals, a 28.0% rise compared to the previous year.[10]

These intensive downward knowledge transfers are reinforced by a national performance framework that grades the alliances on clinical integration, providing a clear signal of policy priorities. Specifically, the UMA policy adopted the pre-existing 2018 national evaluation plan, under which lead tertiary hospitals are assessed across five operational dimensions: alliance governance, cross-tier division of labor, vertical resource flows, service efficiency, and long-term institutional sustainability.[11] Although the national plan does not set a uniform value standard for each dimension, the structural distribution of the performance indicators provides a clear signal of central policy intent. Of the eighteen performance criteria employed to evaluate lead hospitals, nearly half explicitly track the downward transfer of clinical expertise, dispatched personnel, and medical technology to member hospitals. By contrast, cost containment is confined to a single criterion, and academic research output is omitted entirely. Clinical integration and downward technical support are therefore functionally central to the evaluation. In turn, this framework gave local health authorities and hospital administrators strong incentives to direct alliance resources toward standardizing clinical practices and improving service quality at member hospitals.

3. Data sources, data construction, and sample overview

3.1. Data sources

To provide causal evidence on the trade-off between exploitation and exploration, our empirical analysis combines two main data sources. We measure exploitation through a novel application of LLM to patient-level medical records, which allows us to construct a granular measure of clinical practice alignment. We measure exploration by tracking physicians' academic research output using a comprehensive publication database.

[5] All figures are from the China Health Statistics Yearbook 2021 (National Health Commission of the People's Republic of China, 2021). In 2020, China had 2996 tertiary hospitals out of 35,394 hospitals nationwide, or 8.5% of the total.

[6] The policy was jointly issued by the National Health Commission (NHC) and the National Administration of Traditional Chinese Medicine (NATCM). According to policy documents, the timeline of the UMA policy can be divided into three phases. The first phase, spanning May to June 2019, involved the National Health Commission identifying and selecting pilot cities. In the second phase, from June to July 2019, provincial health administrative departments formulated detailed implementation plans based on the national framework and provided guidance to the selected cities. The third phase, lasting from August 2019 to December 2020, focused on ensuring the smooth implementation of the policy through two evaluations conducted by the National Health Commission. Based on this timeline, we identify June 2019 as the official starting point of the policy. For more details about the policy, please refer to the official document at https://www.gov.cn/zhengce/zhengceku/2019-10/08/content_5437013.htm.

[7] China's healthcare reforms have evolved over time. Initial efforts in 2009 explored collaborations between urban hospitals and community health centers. The 2015 Hierarchical Medical System formalized referral pathways, encouraging non-tertiary facilities to manage chronic conditions (Zhou et al., 2021). The 2019 launch of the UMA coincided with national drug price negotiations designed to lower the costs of listed medicines. Unlike the UMA, which centers on knowledge transfer, this concurrent reform operated strictly through price mechanisms. We empirically examine potential confounding from this and other overlapping policies in Section 7.

[8] The UMA differs from vertical integration models commonly seen in the United States. While vertical integration in the U.S. typically focuses on merging healthcare providers and insurers to streamline costs and improve care coordination within a single entity, the UMA policy emphasizes collaboration and resource sharing between independent institutions at different tiers of care. This approach prioritizes enhancing the capabilities of lower-tier hospitals through mentorship, training, and knowledge transfer from tertiary hospitals, without necessarily consolidating administrative or financial structures (Ye et al., 2019).

[9] Detailed government policy documents regarding the UMA policy, including selection criteria and implementation guidelines, can be found at https://www.gov.cn/zhengce/zhengceku/2019-10/08/content_5437013.htm. Because these criteria are closely linked to baseline administrative and healthcare infrastructure, our robustness analysis in Section 7 accounts for differential city-level trends in economic conditions and healthcare supply.

[10] Data sourced from official government reports. For more details, please refer to https://www.gov.cn/zhengce/2020-07/31/content_5531670.htm.

[11] The 2018 *Medical Consortium Comprehensive Performance Evaluation Work Plan (Trial)* was jointly issued by the National Health Commission and the National Administration of Traditional Chinese Medicine. This document established the national standard for assessing the operational success of all medical alliances, including the subsequent UMA pilots. Moreover, the linkage between evaluation results with fiscal subsidies, insurance reimbursements, and compensation budgets appears explicitly in the UMA pilot notice.

Measuring Exploitation: Clinical Practice Data. Our primary data source for measuring clinical practice alignment—our proxy for exploitation—is the China National Clinical Research Center for Neurological Diseases (NCRC-ND), a national platform that aggregates high-quality, detailed patient records from a consortium of hospitals.[12] Our raw dataset spans from 2015 to 2020 and includes records for 310,447 patients from 30 hospitals (17 tertiary and 13 secondary) located across 27 cities. To ensure a clean and clinically meaningful comparison of practice styles, we limit our sample to patients diagnosed with cerebrovascular conditions, which constitute the primary clinical focus of the NCRC-ND network.[13]

A key element of our research design is to measure alignment against an established standard of care. To do so, we focus on three key components of the NCRC-ND database: patient admission records, radiology examination reports, and physician medication orders. The admission records provide essential patient demographics, insurance status, and cost data. The radiology reports, sourced from a subset of 15 hospitals (nine tertiary and six secondary), offer detailed, unstructured textual interpretations of neuroimaging examinations. This rich textual data is the foundation of our methodology for identifying clinically analogous cases, as elaborated in Section 3.2.1. Finally, the medication orders provide structured data on prescribed treatments, allowing us to quantify and compare clinical decisions.

To clarify the composition of the analytical samples, we report sample sizes separately for the practice-alignment and downstream-outcome analyses. The practice-alignment analysis uses 34,145 patient-level observations from the six secondary hospitals with both radiology reports and medication-order information. The downstream analyses draw on structured records from all 13 secondary hospitals: the inpatient-duration and imaging-utilization samples contain 83,220 observations, while the expenditure samples are smaller because of outcome-specific missing data. Since each city in our dataset typically corresponds to a single hospital, our analysis naturally lends itself to an intent-to-treat (ITT) framework. We assign treatment status according to the official UMA pilot roster. A hospital is classified as treated if it is located in a designated pilot city or a designated pilot district within a provincial-level municipality.

Measuring Exploration: Academic Innovation Data. To measure physicians' engagement in exploratory research, we collect data on academic publications from the China National Knowledge Infrastructure (CNKI), a comprehensive database of Chinese scholarly articles analogous to Web of Science or Google Scholar.[14] For each article, CNKI provides detailed metadata, including author affiliations, journal information, publication date, and citation counts, which are essential for assessing both the quantity and quality of research output.

We construct our hospital-level sample for the exploration analysis as follows. First, we identify all secondary hospitals located in the same local administrative units as the hospitals contributing radiology examination reports to our clinical dataset. This approach ensures that the hospitals in our exploitation and exploration analyses are drawn from the same geographic and policy environments. Because Beijing enters the design only as the location of the benchmark hospital, supplying the standard of care rather than a study subject, we exclude its secondary hospitals from the publication sample and keep the two analyses on the same set of cities. Our final sample frame for this analysis consists of 653 secondary hospitals located across 49 local administrative units.[15] We then collect all journal articles published by these hospitals and indexed in CNKI between January 2015 and December 2022. An article is attributed to a hospital if at least one of its authors lists that hospital as their affiliation. Finally, to focus on research-active institutions, we exclude hospitals that have zero publications over the entire sample period. This process yields our final panel for analysis. Our final analytical sample therefore consists of a balanced panel of 541 secondary hospitals, yielding a total of 51,936 hospital-month observations. For the across-tier placebo tests, we also construct a tertiary-hospital publication panel, containing 198 tertiary hospitals and 19,008 hospital-month observations.

3.2. Key variables

3.2.1. The measure of practice style alignment: A large language model approach

A central challenge in measuring exploitation is to develop an objective, granular measure of how closely a physician's practice style aligns with a "best-practice" benchmark. Standard approaches that rely on structured administrative data, such as International Classification of Diseases (ICD) codes, are often too coarse for this task. They can identify a patient's general diagnosis but fail to capture the rich clinical nuance—such as the specific location, severity, and characteristics of a condition—that is critical for an "apples-to-apples" comparison of treatment decisions. These limitations motivate our use of unstructured clinical text from radiology reports, which contain precisely this detailed diagnostic information (Gentzkow et al., 2019).

However, the unstructured narrative format of these reports presents its own analytical complexities. Traditional text analysis methods, such as those based on word counts or dictionaries (i.e., "bag-of-words"), are inadequate because they ignore the sequential and contextual nature of language (Loughran and McDonald, 2016). For example, in our setting, such methods could easily confuse a report stating "the hematoma shows significant growth" with one noting "the hematoma shows significant resolution". While both sentences share keywords like "hematoma" and "significant", their clinical meanings are polar opposites: the former suggests a worsening condition that may require urgent intervention, while the latter indicates recovery. To overcome these measurement challenges, we use a large-language-model (LLM) encoder to represent each radiology report in semantic space. Pre-trained on large textual corpora, LLM-based encoders can capture contextual meaning beyond shared keywords, making them well suited for distinguishing clinically different statements that appear lexically similar (Devlin et al., 2019; Reimers and Gurevych, 2019; Gao et al., 2021). This capability allows us to transform the complex clinical narrative of each report into a high-dimensional vector embedding—a numerical representation that captures its precise clinical meaning.

Specifically, our methodology for constructing the practice style alignment measure proceeds in three stages. The first, and most critical, involves using an LLM to quantify the textual similarity between clinical cases. The success of this matching hinges on generating high-quality vector embeddings that are sensitive to semantic meaning. For this purpose, we employ SimCSE (Simple Contrastive Learning of

[12] The NCRC-ND, affiliated with Beijing Tiantan Hospital, Capital Medical University, was established in 2013 as the first officially recognized clinical research center for neurological diseases in China, approved by the Ministry of Science and Technology and the National Health Commission. For more information, see the center's official website: https://ncrcnd.org.cn/Center_Positioning/.

[13] Specifically, cerebrovascular conditions were identified using ICD-10 three-digit codes: I60–I69, G45, and G46. This selection retains 68.19% of the original patient records.

[14] CNKI (China National Knowledge Infrastructure) is a national-level academic database established through initiatives by key government agencies, including the Ministry of Education and the Ministry of Science and Technology.

[15] The sample spans 12 city-level areas: 11 prefecture-level cities and the municipality of Chongqing. For Chongqing, the official UMA pilot roster assigns treatment at the district level, offering us more granular information for tracking the treatment status of hospitals. The 49 local administrative units therefore consist of 11 prefecture-level cities outside Chongqing and 38 districts within Chongqing. The publication results are robust to adding Beijing secondary hospitals and are available upon request.

Sentence Embeddings), a contrastive-learning method that fine-tunes a pre-trained language-model encoder (e.g., BERT) to produce embeddings suitable for semantic similarity comparison (Devlin et al., 2019; Gao et al., 2021).

The model is fine-tuned on radiology reports from our sample using an unsupervised contrastive learning framework. This approach is powerful as it requires no pre-labeled data. Instead, it generates its own training pairs: a single radiology report is passed through the model twice, each time with a different random "dropout" pattern applied. This process yields two distinct embeddings that are, by construction, semantically identical. The model is then trained to treat these as a "positive pair", minimizing the distance between their vector representations. By repeating this process across the training corpus, the LLM adapts its general linguistic knowledge to the specific syntax and terminology of our clinical domain. The resulting fine-tuned model transforms each report into a high-quality vector embedding.[16] For each patient case in a secondary hospital, we then compute the cosine similarity between its report's vector and all report vectors from our benchmark hospital, identifying the benchmark case with the highest similarity as the "clinically analogous match".[17]

Second, for all identified matched case pairs, we codify their respective prescription patterns into a comparable format. We classify all prescribed medications into 32 distinct categories based on China's National Essential Medicines List (2018 edition), a standardized government framework designed to guide clinical practice.[18] This process transforms each patient's prescription into a 32-dimensional binary vector, where an element equals one if a drug from that category was prescribed and zero otherwise.

Lastly, we calculated the similarity of the medical prescriptions for a specific case from the test hospital (Case A) and the prescriptions for its most similar counterpart from the benchmark hospital (Case A′). The similarity score for the prescriptions associated with Case A is computed using the formula:

$$\text{esim}_{A,A'} = 1 - \frac{len\left(set(A)\Delta set(A')\right)}{32} \tag{1}$$

where $set(A)$ denotes the set of medications prescribed in Case A, and $set(A')$ denotes the set of medications prescribed in Case A'. The operation Δ calculates the symmetric difference between the two sets, defined as the union of both sets minus their intersection, reflecting the number of unique elements in the two sets. The function $len()$ counts the number of elements in the resulting set. If the two medication sets share many common medications, the symmetric difference will have fewer elements, indicating a higher similarity between the cases. Conversely, if the medication types in the two sets differ significantly, the symmetric difference will contain more elements, suggesting a lower similarity. To standardize the result, we divide the number of elements in the symmetric difference by the total number of medication categories (i.e., 32), allowing the similarity score to range from 0 to 1. For easier interpretation, we subtract the normalized result from 1, ensuring that a higher value of $\text{esim}_{A,A'}$ reflects greater similarity in the prescribed medications.

3.2.2. The measure of physician innovation output

To capture the exploration dimension of the trade-off, our second key task is to construct a meaningful measure of physician innovation. We use academic publications as our primary proxy, as they are particularly well-suited for capturing the conceptual and clinical advances by practicing physicians, which are often not patentable but are fundamental to medical progress (Adams, 1990; Murray and Stern, 2007; Aghion et al., 2008). However, just as administrative codes offer a coarse view of clinical practice, a simple count of publications can be a noisy and potentially misleading measure of genuine innovation. In an environment shaped by "publish or perish" pressures, not all publications represent significant exploratory effort; many may be routine or incremental outputs designed to meet institutional requirements. We therefore adopt a multi-faceted measurement strategy that assesses not only the quantity of research but also its quality and subsequent impact. As a baseline measure of quantity, we track the total number of articles published by physicians affiliated with each hospital. We refine this by also focusing on first-author publications, which, following conventions in medical research, more precisely capture the innovative efforts originating from within the hospital rather than peripheral contributions to larger projects led elsewhere (Agha and Molitor, 2018).

Beyond simple counts, a more robust measurement strategy must account for the quality and subsequent impact of the research. To differentiate significant research from more routine work, we measure quality by tracking publications in top-tier journals. Defining "top-tier" in the context of Chinese-language journals requires careful consideration; we classify journals with an impact factor greater than 1 as top-tier, a meaningful threshold given that the average impact factor for comprehensive Chinese medical journals is substantially lower.[19] This metric allows us to proxy for research that has successfully passed a more rigorous peer-review process. Finally, to capture the long-run influence, we track the cumulative citation counts for each hospital's publications. Citations serve as a market-based signal of a paper's value to the scientific community, providing a dynamic, post-publication measure of its impact (Jaffe et al., 1993; Azoulay et al., 2011). For all these metrics, we exclude non-scholarly publication types such as newspapers and industry standards to focus purely on peer-reviewed knowledge creation.

3.3. Data overview

Table 1 presents summary statistics for patient-level outcomes and hospital-level research productivity in secondary hospitals, separately for the pre- and post-policy periods. Our measure of prescription-practice alignment relies on radiology report data available for six

[16] To make the measurement procedure transparent and reproducible, Appendix B documents the full language-model pipeline, including the base encoder, fine-tuning corpus, preprocessing steps, train–test separation, hyperparameters, and validation exercises.

[17] We use cosine similarity because it measures the orientation (i.e., angle) between two vectors, rather than their magnitude. In the context of text embeddings, the vector's direction encodes its semantic meaning, while its length is often influenced by factors like sentence length and may not be directly interpretable. Two reports describing the same clinical finding should have vectors pointing in a similar direction, even if their lengths differ. Cosine similarity captures this semantic alignment while being robust to differences in vector magnitude, which makes it the standard and most appropriate metric for this task.

[18] The National Essential Medicines List (2018 edition) is an official catalog of essential medicines published by the Chinese government. The list includes a carefully selected range of medicines that are considered essential for addressing the most common healthcare needs in China. It aims to guide the procurement, distribution, and prescription of medicines in healthcare facilities across the country, helping to standardize treatment protocols and ensuring that essential medicines are consistently available in both urban and rural healthcare settings. Under the national essential-medicines system introduced in 2009, the list was revised in 2012 and again in 2018. The 2018 edition took effect before the implementation of the UMA policy and remained unchanged throughout our post-policy sample period. In our analysis, we use this edition solely as a fixed classification framework to group prescribed medications into 32 categories, applying the same mapping consistently across all hospitals and periods.

[19] The average impact factor for comprehensive Chinese medical journals in 2019 was only 0.606, making an impact factor of 1 a meaningful, high-quality threshold. This benchmark is frequently used in university and hospital promotion systems to identify significant research contributions. For more details, see https://blog.sciencenet.cn/blog-475824-1222991.html.

Table 1
Summary statistics.

Variable	Pre-Policy				Post-Policy			
	Mean (1)	Median (2)	Std. dev. (3)	Obs (4)	Mean (5)	Median (6)	Std. dev. (7)	Obs (8)
Panel A: Patient-level outcomes and characteristics (Main regression sample)								
Key outcomes								
Similarity score	0.877	0.888	0.070	24,756	0.899	0.912	0.070	9389
Inpatient duration (Days)	12.565	13.000	8.698	24,756	11.001	10.000	8.021	9389
Radiological exam incidence (%)	0.830	1.000	0.375	24,756	0.849	1.000	0.358	9389
Number of radiological exams	1.179	1.000	0.909	24,756	1.270	1.000	0.984	9389
Total expenditure (RMB)	7819.802	5460.580	15,666.017	23,083	7188.916	4763.310	11,734.514	9387
Drug expenditure (RMB)	3226.468	2296.690	3685.700	20,893	2795.260	1809.430	3513.899	8118
Non-drug expenditure (RMB)	4806.129	3171.030	14,345.296	20,893	4667.562	3006.730	10,231.417	8118
Characteristics of patients								
Age (Years)	65.196	66.000	11.583	24,756	65.420	66.000	11.793	9389
Male (%)	0.504	1.000	0.500	24,756	0.520	1.000	0.500	9389
Married (%)	0.648	1.000	0.478	24,756	0.629	1.000	0.483	9389
Insurance type (%)								
Urban Employee	0.153	0.000	0.360	24,756	0.101	0.000	0.301	9389
Urban Resident	0.358	0.000	0.480	24,756	0.441	0.000	0.496	9389
New Rural Cooperative	0.386	0.000	0.487	24,756	0.368	0.000	0.482	9389
Poverty Medical Aid	0.004	0.000	0.066	24,756	0.009	0.000	0.094	9389
Private Commercial	0.000	0.000	0.006	24,756	0.000	0.000	0.000	9389
Public Funding	0.030	0.000	0.171	24,756	0.016	0.000	0.125	9389
Out-of-Pocket	0.064	0.000	0.245	24,756	0.055	0.000	0.227	9389
Other	0.004	0.000	0.063	24,756	0.011	0.000	0.106	9389
Panel B: Hospital-level research productivity								
Monthly research output by hospitals								
Total	0.826	0.000	1.854	28,673	0.631	0.000	1.637	23,263
First-Author	0.755	0.000	1.755	28,673	0.564	0.000	1.530	23,263
Top-Journal	0.213	0.000	0.668	28,673	0.136	0.000	0.547	23,263
Citations	4.101	0.000	11.128	28,673	2.198	0.000	6.798	23,263

Note: This table presents the mean, median, and standard deviation for each variable, with the sample divided into pre- and post-policy periods (i.e., before and after June 2019). Panel A displays patient-level characteristics from the six secondary hospitals with available radiological records. Insurance categories are defined as follows: *Urban Employee* (Urban Employee Basic Medical Insurance), *Urban Resident* (Urban Resident Basic Medical Insurance), *New Rural Cooperative* (New Rural Cooperative Medical Scheme), *Poverty Medical Aid* (medical poverty assistance), *Private Commercial* (commercial health insurance), *Public Funding* (fully publicly funded), *Out-of-pocket* (fully self-paid), and *Other* (other insurance types). Panel B presents statistics for 541 secondary hospitals located within the same city as our sample of secondary hospitals.

secondary hospitals, yielding a patient-level sample of 34,145 observations. Patients in this sample are approximately 65 years old on average, the gender composition is roughly balanced, and nearly two-thirds are married. The share of patients classified as fully paying out of pocket declines from 6.4% in the pre-policy period to 5.5% in the post-policy period.

Regarding the alignment of prescription styles, the unconditional mean prescription similarity score increases from 0.877 before the policy to 0.899 afterward, indicating greater average alignment with the benchmark hospital in the post-policy period. Inpatient length of stay exhibited a reduction from 12.57 to 11.00 days, potentially reflecting efficiency gains resulting from enhanced care coordination. Both the incidence of radiological examinations and the number of examinations per patient increase, from 83.0% to 84.9% and from 1.18 to 1.27, respectively. Average total medical expenditure declines from RMB 7819.8 to RMB 7188.9, while average drug expenditure falls from RMB 3226.5 to RMB 2795.3. These descriptive patterns provide an initial view of how outcomes evolved around the policy, while the difference-in-differences analysis below distinguishes the UMA effect from broader contemporaneous changes.

Panel B presents hospital-month research output for secondary hospitals. All four measures are lower in the post-policy period. Specifically, total publications fell from 0.826 to 0.631 per hospital-month, first-author publications dropped from 0.755 to 0.564, top-journal publications decreased from 0.213 to 0.136, and total citations plummeted from 4.10 to 2.20. These findings suggest that while certain secondary hospitals may benefit from collaborative opportunities within medical alliances, the overall research capacity of secondary hospitals appears to have diminished, potentially reflecting a trade-off between the standardization of clinical practices and local research innovation.

3.4. Descriptive statistics

Prior to conducting causal analysis, we begin by presenting descriptive evidence that motivates our study: the existence of pre-policy disparities in clinical practice styles. We compare the prescription patterns of secondary and tertiary hospitals against our expert benchmark hospital, focusing on the period before the UMA policy's implementation.

Fig. 1 plots the kernel density of deviations in prescription frequencies across 32 medication categories relative to the benchmark hospital. The distribution for tertiary hospitals is tightly centered around zero, suggesting that their prescribing patterns are broadly aligned with the benchmark. In comparison, the distribution for secondary hospitals is more dispersed and exhibits a heavier upper tail. Although the two tiers have similar central tendencies, secondary hospitals show a higher mean absolute deviation and a larger share of observations with substantial deviations, as indicated by the summary statistics reported in the figure. Figure A2 reinforces this pattern by plotting the cumulative distribution function of the absolute deviations. The steeper curve for tertiary hospitals confirms a higher concentration of small deviations, whereas the rightward shift in the secondary hospitals' curve reflects a greater frequency of large deviations from the benchmark. Together, these baseline differences suggest that secondary hospitals had more room to adjust toward benchmark practices, providing descriptive motivation for studying whether the UMA promoted knowledge diffusion and practice convergence within medical alliances.

To explore potential policy-induced shifts, we next examine changes in prescription probabilities across the 32 categories for secondary hospitals before and after the UMA policy, relative to the benchmark hospital. Fig. 2 visualizes these changes, with categories ordered by

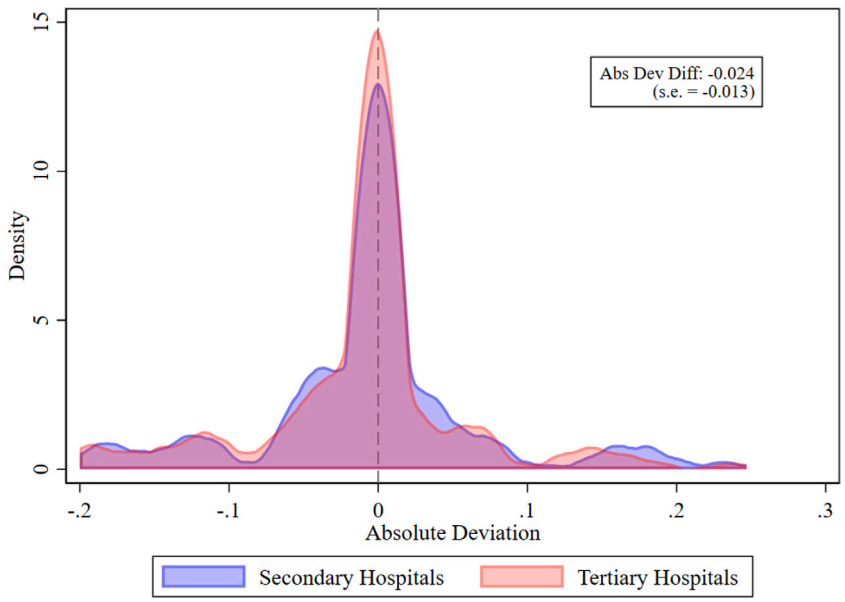


Fig. 1. Density of deviations in prescription frequencies across secondary and tertiary hospitals.
Note: This figure illustrates the density distribution of deviations in prescription frequencies across drug categories for secondary and tertiary hospitals relative to the benchmark hospital. The x-axis measures the deviation in prescription frequency for each drug category, defined as the difference between the frequency in a secondary or tertiary hospital and that in the benchmark hospital. The sample comprises 32 drug categories across 6 secondary and 8 tertiary hospitals, yielding 448 deviations (192 for secondary, 256 for tertiary). The annotation in the upper-right corner reports the difference in the mean of the absolute values of these deviations between secondary and tertiary hospitals (mean for secondary hospitals minus mean for tertiary hospitals), along with the standard error, based on a two-sample t-test. Furthermore, the 90th-percentile absolute deviation is 0.461 for secondary hospitals versus 0.298 for tertiary hospitals, reflecting the heavier tail in the secondary distribution. The blue line represents the density for secondary hospitals, and the red line represents the density for tertiary hospitals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the benchmark hospital's usage frequency (highest to lowest) and the bottom 16 aggregated as "Others". Post-policy, we observe declines in prescription probabilities for most categories—excluding anti-inflammatory drugs and neurological drugs—particularly in symptom-control and palliative care medications like dermatological drugs. These shifts point to a broad contraction in prescribing after the UMA policy. Although this pattern does not by itself establish convergence, given the substantial variation in benchmark usage across categories, it is consistent with more restrained and selective prescribing. The figure therefore provides suggestive evidence of improved prescribing discipline while offering mixed evidence on convergence, which we assess more formally using the high-dimensional alignment outcome and the vector-distance evidence presented in Section 5.

4. Empirical strategy

4.1. Empirical strategy

Under the assumption of parallel trends, the implementation of the UMA policy provides a unique opportunity to estimate its causal effect using a difference-in-differences (DID) strategy. Leveraging the simultaneous implementation of the policy across designated pilot areas as a natural experiment, we compare changes in outcomes pre- and post-policy between hospitals in cities where the UMA policy was implemented (treated group) and those in cities without the policy (control group). The parallel trends assumption posits that, in the absence of the UMA policy, the changes in physician practice styles and academic innovation outcomes between treated and control hospitals would have remained constant over time. This assumption allows us to isolate the effect of the UMA policy by attributing any observed divergence in trends between the two groups to the policy's implementation.

First, to identify the causal impact of knowledge transfer facilitated through the UMA policy on physician prescription behaviors, we estimate the following regression equation using patient-level data:

$$Y_{idht} = \delta \,\mathrm{Post}_t \times \mathrm{Treat}_h + \beta \mathbf{X}_{idht} + \tau_{dh} + \tau_t + \varepsilon_{idht} \tag{2}$$

where Y_{idht} denotes the outcome variable of interest for patient i with disease d, admitted to hospital h during period t (measured monthly). Outcomes include the similarity score of prescriptions (based on matched radiological reports), medical expenditures, and healthcare utilization metrics. Post_t is a dummy variable that equals 1 if the admission occurred on or after June 2019, the start of the UMA policy. Treat_h is a dummy variable that equals 1 if hospital h is located in a treated city.[20] The coefficient δ captures the intent-to-treat effect of the UMA policy, reflecting both direct knowledge spillovers (via formal alliances) and indirect spillovers (through informal physician networks).

The vector $\mathbf{X}_{idht}$ includes patient-level controls: age, gender, marital status, insurance type, and the treating physician's experience (proxied by pre-policy patient volume). We also include city-level monthly COVID-19 case counts to account for pandemic-related disruptions that may asymmetrically affect treatment and control groups. τ_{dh} represents hospital–disease fixed effects, which capture unobserved, time-invariant heterogeneity at the hospital–disease level, such as differences in treatment protocols for specific diseases across hospitals.[21] Year–month fixed effects, τ_t, capture time-varying national-level shocks

[20] We use each secondary hospital's recorded address to assign treatment status according to the official UMA pilot roster.

[21] As found in the literature, hospitals may adopt different treatment approaches for various diseases due to their comparative advantages in certain specialties (Chandra and Staiger, 2020). The hospital–disease fixed effects in this specification allow us to account for these comparative advantages, offering a more precise control for hospital-specific treatment norms across different diseases. For robustness, we also report results from specifications where hospital–disease fixed effects, τ_{dh}, are replaced by hospital fixed effects and disease fixed effects, τ_d and τ_h, respectively.

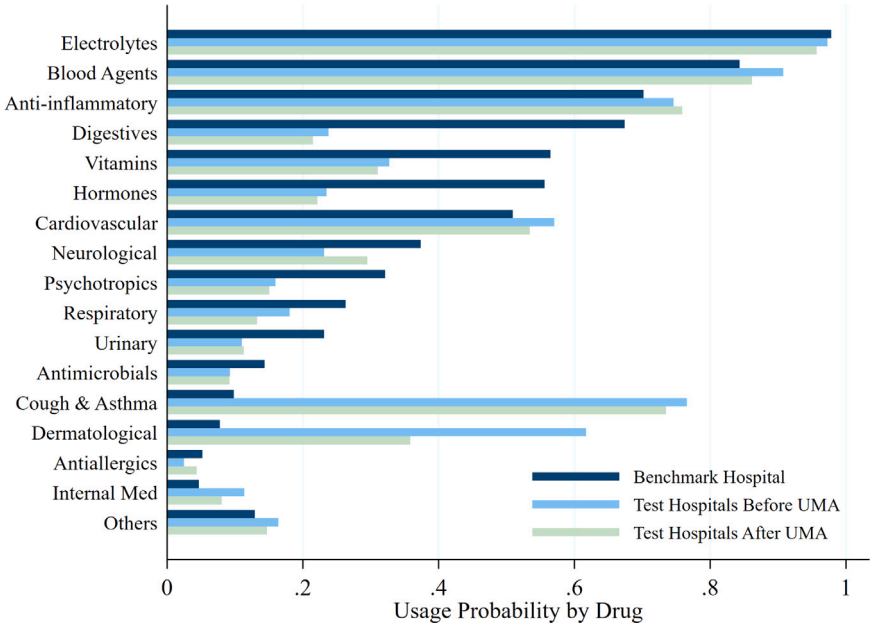


Fig. 2. Usage probabilities of drug categories in secondary hospitals before and after UMA policy.
Note: This figure illustrates the usage probabilities of drug categories in secondary hospitals before and after the UMA policy, relative to a benchmark hospital. The *x*-axis measures the probability of prescribing a drug within a given category. The *y*-axis lists 32 drug categories, with the bottom 16 categories aggregated into an "Others" group, ordered by the benchmark hospital's usage probabilities (from highest to lowest). Dark blue bars represent the benchmark hospital's usage probabilities, light blue bars represent secondary hospitals' usage probabilities before the UMA policy, and light green bars represent secondary hospitals' usage probabilities after the UMA policy. The sample includes 6 secondary hospitals, with usage probabilities calculated across all prescriptions in each period. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that could affect the outcomes.[22] Standard errors are clustered at the hospital–disease level.[23]

There may be concerns regarding the validity of the parallel trends assumption, particularly the possibility that cities implementing the policy could exhibit different trends in knowledge spillover or physician prescription behaviors prior to its implementation. To address this, we estimated a fully dynamic version of Eq. (2) to check for potential pre-trends. The event study regression is specified as follows:

$$Y_{idht} = \delta_k \times \sum_{k\neq -1} \mathbf{1}\{t - t_h = k\} \times \text{Treat}_h + \beta \mathbf{X}_{idht} + \tau_{dh} + \tau_t + \varepsilon_{idht} \quad (3)$$

where $\mathbf{1}\{t - t_h = k\}$ refers to a set of indicator variables assigned a value of one if the patient admitted to hospital h in period t is k months away from the implementation of the policy (t_h). The month just prior to the policy implementation is treated as the omitted category, and comparisons are made with those exposed to the policy in other relative months. All other specifications align with those outlined in Eq. (2).

Second, to estimate the causal impact of the UMA policy on academic innovation outcomes, we use hospital-level data and estimate the following regression equation:

$$\log(Y_{hct}) = \gamma\, \text{Post}_t \times \text{Treat}_c + \eta_h + \eta_t + \xi_c \times t + \varepsilon_{hct} \quad (4)$$

where $\log(Y_{hct})$ represents the dependent variables: (i) the logarithm of one plus the total number of publications for hospital h in city c during period t, and (ii) the logarithm of one plus total citations for hospital h in city c during period t.[24] Post_t is a dummy variable equal to 1 for periods on or after the implementation of the UMA policy (June 2019), while Treat_c equals 1 if the hospital is located in a treated city. η_h denotes hospital fixed effects, and η_t represents the year–month fixed effects. One potential concern is that the number or quality of publications could correlate with broader city-level trends that independently influence academic outcomes. To address this, we include $\xi_c \times t$ to account for city-level time trends. Standard errors are clustered at the local administrative unit level, corresponding to the level of policy assignment.

Similarly, to examine the validity of the parallel trends assumption, we estimated a fully dynamic version of Eq. (4) as follows:

$$\log(Y_{hct}) = \gamma_g \times \sum_{g=-5,\, g\neq -1}^{3} \mathbf{1}\{t - t_c = g\} \times \text{Treat}_c + \beta \mathbf{X}_{ct} + \eta_h + \eta_t + \xi_c \times t + \varepsilon_{hct} \quad (5)$$

where $\mathbf{1}\{t - t_c = g\}$ denotes a set of event-study dummies that equal one when hospital outcomes in city c at time t are observed g years relative to policy implementation at t_c. The year just prior to the policy implementation is treated as the omitted category, and comparisons are made with those exposed to the policy in other relative years. All other specifications align with those outlined in Eq. (4).

A notable concern arises from the use of $\log(y + 1)$ transformations, which can introduce unit dependence by conflating two distinct margins of treatment effects: the extensive margin (the likelihood of transitioning from zero to a positive outcome) and the intensive margin (the magnitude of change among positive outcomes) (Mullahy and Norton, 2022; Chen and Roth, 2024). Different scaling of y may

[22] As we account for year–month fixed effects, which offer more granularity than year-level city controls, we opted not to include year-level city controls.

[23] Our preferred specification includes hospital–disease fixed effects, year–month fixed effects, patient-level controls, and the city-month COVID-19 confirmed-case count. In Section 7, we also report progressively richer specifications that begin with separate hospital and disease fixed effects, replace them with hospital–disease fixed effects, and then add patient-level and COVID-19 controls.

[24] As discussed in Section 3, the publication sample covers 11 prefecture-level cities, as well as 38 districts within Chongqing, at which the treatment status is defined.

Table 2
DID estimates of the policy's effect on prescription similarity.

	Dependent variable: Esim score			
	(1)	(2)	(3)	(4)
Post × Treat	0.0236***	0.0215***	0.0201***	0.0202***
	(0.0025)	(0.0022)	(0.0026)	(0.0026)
Year–Month FE	✓	✓	✓	✓
Hosp. FE	✓			
Disease FE	✓			
Hosp-Disease FE		✓	✓	✓
Individual controls			✓	✓
COVID-19 control				✓
No. of Obs.	34,145	34,145	34,145	34,145
R^2	0.108	0.116	0.126	0.126

Note: This table reports the intent-to-treat estimates of the UMA policy implementation on physicians' practice styles in secondary hospitals. Specifically, it presents estimates of the coefficient δ from Eq. (2), where the dependent variable measures the alignment of physicians' practices based on the similarity of prescriptions derived from matched radiological text reports. The detailed similarity score was computed using the methodology outlined in Section 3.2.1. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 3
DID estimates of the policy's effect on prescription overlap and union counts.

	Overlap count		Union count	
	(1)	(2)	(3)	(4)
Post × Treat	0.4305***	0.4319***	−0.1855**	−0.1878**
	(0.0391)	(0.0409)	(0.0789)	(0.0818)
Pre-Mean	4.390	4.390	8.298	8.298
Year–Month FE	✓	✓	✓	✓
Hosp-Disease FE	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓
COVID-19 control		✓		✓
No. of Obs.	34 145	34 145	34 145	34 145
R^2	0.16	0.16	0.20	0.20

Note: This table reports the intent-to-treat estimates of the UMA policy's effect on prescription patterns in secondary hospitals, focusing on the coefficient δ from the difference-in-differences specification in Eq. (2). The dependent variable in columns (1)–(2) is the overlap count of prescribed medications between the secondary and benchmark tertiary hospitals. Columns (3)–(4) use the union count of prescribed medications as the dependent variable. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

alter the relative weighting of these margins, potentially biasing the interpretation of results. To address this, we report separate results for the extensive and intensive margins in Section 5. Following Chen and Roth (2024), we also estimate policy effects using Poisson regression and linear regression on untransformed outcomes, providing robust alternatives to the log-transformed models. These findings are detailed in Sections 5 and 7.

5. Results

5.1. Main results

5.1.1. Prescription similarity

To causally evaluate the UMA policy's effect on prescription alignment, we employ a DID framework, as outlined in Eq. (2). Our outcome variable is the prescription similarity score, which quantifies the alignment between secondary hospitals' prescriptions and those of the benchmark hospital, derived from matched radiology-prescription pairs (see Section 3.2.1 for details).

Table 2 reports the DID estimates across four specifications. Column (1) includes hospital, disease, and time fixed effects. Column (2) replaces hospital and disease fixed effects with hospital–disease fixed effects for greater granularity. Column (3) adds patient-level controls (age, gender, marital status, insurance type, and treating physician experience), while Column (4) further accounts for COVID-19 impacts. Across all models, the DID coefficient is stable at approximately 0.02 and significant at the 1% level, implying a policy-induced increase in similarity scores of approximately 0.29 standard deviations. These results indicate that policy-induced knowledge spillovers facilitated the alignment of clinical practices in secondary hospitals with the standardized protocols of higher-tier hospitals.

To assess the dynamics of this effect, we estimate an event-study specification (Eq. (3)), presented in Fig. 3. Pre-policy coefficients are statistically indistinguishable from zero, supporting the parallel trends assumption.[25] Following the policy's implementation, practice similarity scores exhibit a significant upward trend, growing to a substantial increase of approximately 0.46 standard deviation by the sixth month. Beginning in the seventh month, coinciding with the onset of COVID-19 lockdowns, the estimates become more volatile, yet the point estimates themselves remain at a high level, consistently comparable to their pre-lockdown peak. This trajectory demonstrates that the policy induced not only a rapid convergence in clinical practices but also that this alignment was sustained throughout the post-implementation period.[26]

Our analysis thus far has established that the UMA policy causally increased prescription similarity scores. We now turn to decomposing this effect to understand its underlying components. The similarity score is a function of two key elements: the union (all medications prescribed by either the test or benchmark hospital) and the intersection (medications prescribed by both). An increase in similarity can arise from either a reduction in the union (fewer unique medications) or an expansion in the intersection (more shared prescriptions).

The evidence shows that similarity increases through both margins, and both move in the direction of greater alignment. As shown in Table 3, the union contracts while the intersection expands. The UMA policy reduced the number of medication types prescribed by either secondary hospitals or the benchmark tertiary hospital by 0.19, a 2% decline relative to the average union count prior to the policy's implementation. Over the same period, the intersection—medications prescribed by both—grew by 0.43, close to a 10% increase relative to its pre-policy mean. Higher similarity therefore reflects two complementary movements: secondary hospitals shed medications outside the benchmark's formulary and, at the same time, took up medications the benchmark already prescribed. The event-study plot in Figure A4 further illustrates this pattern: both components are flat and insignificant before the policy, and from the first post-policy month the intersection rises steadily while the union declines.

Taken together, these movements indicate convergence toward the benchmark rather than a mechanical expansion or contraction of prescribing. The decomposition shows that secondary hospitals both reduced medication categories not used by the benchmark and increased categories already prescribed by the benchmark. At the same time, the benchmark hospital's own prescribing portfolio remained stable between UMA implementation and the onset of COVID-19 (Figure A5), which rules out the possibility that the measured alignment reflects the benchmark standard drifting toward secondary hospitals. The increase in similarity therefore comes from secondary hospitals realigning their prescribing portfolios toward a stable authoritative standard: their practice became more focused and standardized, with less idiosyncratic variation relative to the benchmark.

[25] As an additional check, we apply the sensitivity method of Rambachan and Roth (2023). The positive effect remains robust to deviations from parallel trends up to 1.115 times the largest pre-period movement.

[26] Since the clinical panel ends in December 2020, we cannot determine whether this realignment persisted beyond the observed period and therefore do not interpret it as evidence of permanent realignment.

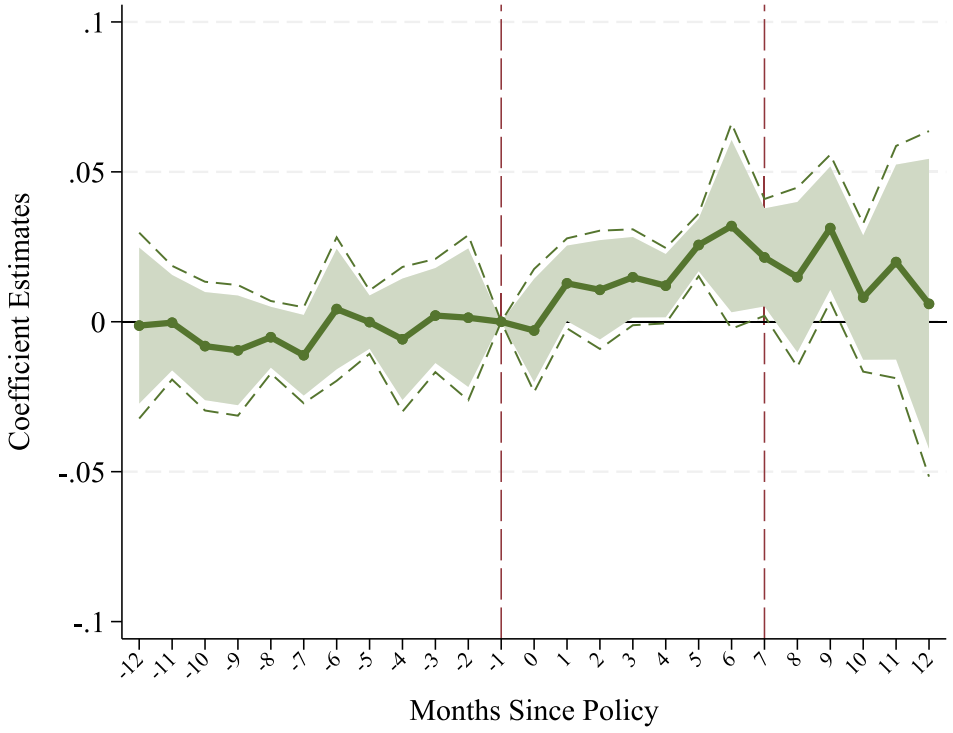


Fig. 3. Event-study estimates of physician prescription similarity.
Note: This figure presents event-study estimates of the dynamic effects of the UMA policy on physician practice alignment in secondary hospitals, covering the period before and after the onset of COVID-19 lockdowns. The dependent variable, prescription similarity relative to benchmark hospital protocols, is constructed via radiology-report matching, as detailed in Section 3.2.1. The left vertical line marks the month before implementation of the UMA policy in June 2019, while the right vertical line indicates the onset of the COVID-19 pandemic in January 2020. The month immediately preceding policy implementation is omitted, normalizing estimates to zero for that period. For clarity and readability, the figure plots only the coefficients for the twelve months on either side of the policy. The earliest pre-policy months rest on very few admissions and are therefore estimated with wide confidence bands that would stretch the vertical scale and flatten the visible movement around the policy date. The full event-study window is reported in Figure A3, where the pre-policy path remains flat and individually insignificant throughout. Dashed lines represent 95 percent confidence intervals, and green shaded areas indicate 90 percent confidence intervals. Standard errors are clustered at the hospital–disease level. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1.2. Academic outcomes

Having established that the UMA policy successfully promoted the exploitation of existing knowledge through practice alignment, we now turn to the other side of the trade-off: its effect on exploration. Table 4 presents estimates of Eq. (4) for academic innovation outcomes. Columns (1)–(4) report changes in the quantity of publications by physicians in secondary hospitals, while columns (5)–(8) focus on changes in the quality of these publications. Odd-numbered columns exclude controls for the COVID-19 pandemic, whereas even-numbered columns additionally account for its impact. The results show that, post-policy, physicians in secondary hospitals experienced a 7% decline in total publications, with first-author publications declining by a larger 7.4%. Publications in top-tier journals decreased by 5.3%, and citation counts fell by 13%. These findings suggest that, in the post-UMA policy period, physician research productivity diminished in both quantity and quality.

As discussed in Section 4, the dependent variables in our DID analysis are expressed as $\log(y+1)$, raising concerns about unit dependence since the treatment effect combines extensive and intensive margins (Mullahy and Norton, 2022; Chen and Roth, 2024). To address this issue, we first examine the distributions of the four outcome variables. As shown in Figure A6, there is no discernible separation between zero and non-zero values across these outcomes. Consequently, the log transformation operates similarly to running a linear regression on untransformed levels of y, primarily capturing the intensive margin effect. We then assess the extensive margin effects, with Table A1 revealing that the UMA policy significantly affects the extensive margin only for publications in top-tier journals, while the effects for the other three outcomes are statistically insignificant. This suggests that the estimates in Table 4 are largely driven by intensive margin dynamics. For further robustness, we follow Chen and Roth (2024) and estimate the policy effects using Poisson regression, finding results consistent with those in Table 4. A detailed discussion of this approach and its implications is provided in Section 7.

Finally, the average treatment effects documented in Table 4 are further validated by event-study estimates based on estimating Eq. (5). As shown in Fig. 4, for each outcome, before the UMA policy, there are no significant differences between treatment and control groups, supporting our assumption of parallel trends. In contrast, in the post-treatment periods, we do see marked declines in the number of publications and citations starting from the year after the UMA policy.

5.2. Downstream outcomes

Our analysis demonstrates that the UMA policy improved prescription alignment with the benchmark hospital while reducing academic productivity. A key question remains: did the policy enhance patient outcomes as intended? Given that healthcare reforms aim to optimize medical resource allocation, we now examine whether the policy-induced changes in physician practices yielded tangible patient-level benefits.

The impact of the UMA policy on patient outcomes could manifest in different ways, depending on how physicians in secondary hospitals respond to knowledge spillovers from tertiary hospitals. On one hand,

Table 4
DID estimates of the policy's effect on academic outcomes.

	Publication quantity				Publication quality			
	Total		First-authored		Top journal		Citations	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post × Treat	−0.0772***	−0.0696***	−0.0820***	−0.0740***	−0.0565***	−0.0532***	−0.1534***	−0.1304***
	(0.0241)	(0.0234)	(0.0247)	(0.0244)	(0.0124)	(0.0111)	(0.0506)	(0.0484)
Hospital FE	✓	✓	✓	✓	✓	✓	✓	✓
Year–Month FE	✓	✓	✓	✓	✓	✓	✓	✓
City linear trends	✓	✓	✓	✓	✓	✓	✓	✓
COVID-19 control		✓		✓		✓		✓
No. of Obs.	51 936	49 464	51 936	49 464	51 936	49 464	51 936	49 464
R^2	0.59	0.59	0.59	0.58	0.32	0.31	0.50	0.50

Note: This table reports the effect of the UMA policy implementation on physicians' academic outcomes in secondary hospitals. Specifically, it presents estimates of the coefficient γ from Eq. (4). The dependent variables are as follows: columns (1)–(2) measure total publications, columns (3)–(4) focus on first-authored publications, columns (5)–(6) represent publications in top-tier journals (i.e., journals with an impact factor greater than 1; see Section 3.2.1 for a discussion of the threshold's rationale), and columns (7)–(8) capture total citations. Odd-numbered columns do not control for the impact of the COVID-19 pandemic, while even-numbered columns include controls for it. Notably, due to the unavailability of COVID-19 case data for the municipality of Chongqing from January 2020 to December 2021, the even-numbered columns include 2472 fewer observations. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

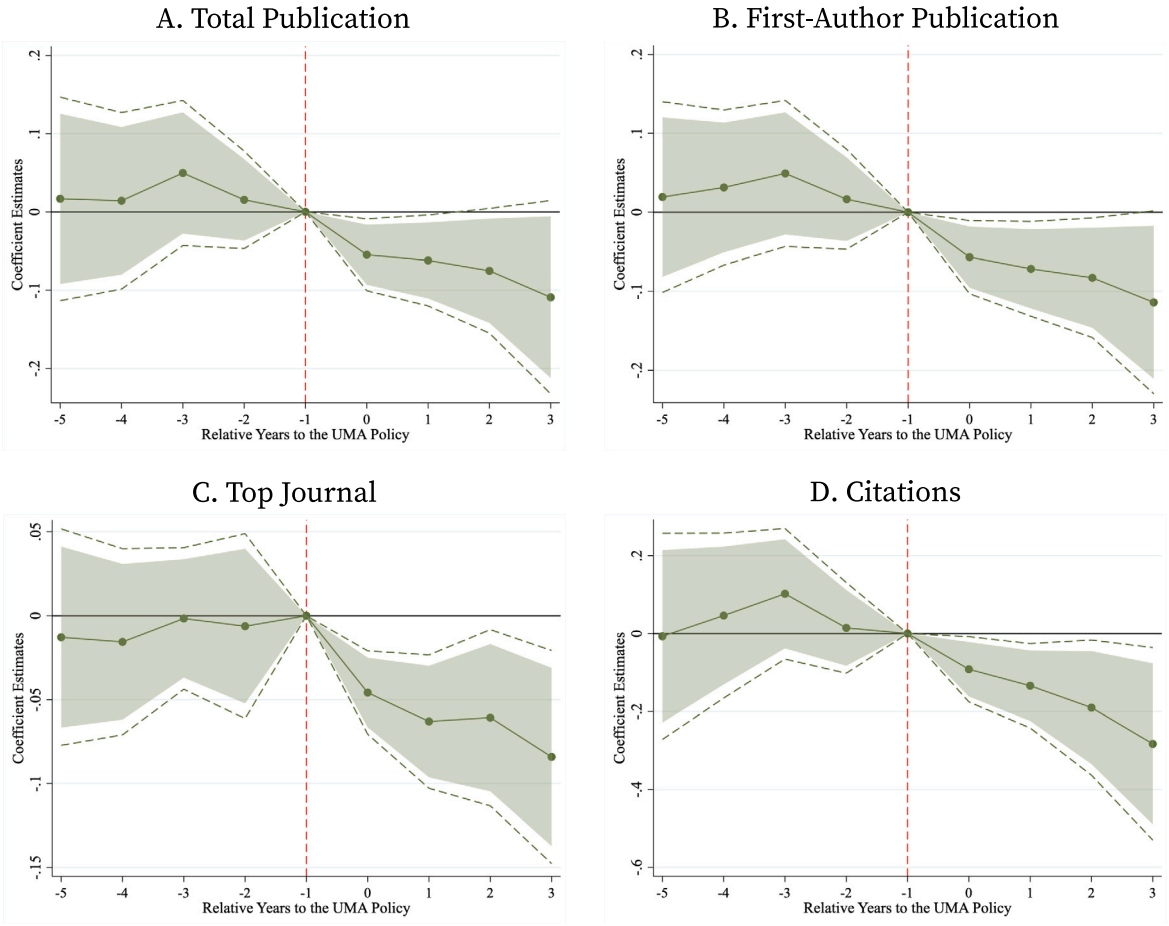


Fig. 4. Event study estimates.
Note: This figure presents event-study estimates of the dynamic effects of the UMA policy implementation on physicians' academic outcomes in secondary hospitals. The dependent variables are as follows: Subfigure (A) measures total publications, Subfigure (B) focuses on first-authored publications, Subfigure (C) represents publications in top-tier journals (i.e., journals with an impact factor greater than 1; see Section 3.2.1 for a discussion of the threshold's rationale), and Subfigure (D) captures total citations. The year immediately preceding the policy implementation is omitted and serves as the reference period, with estimates normalized to zero for that time. Dashed green lines represent 95 percent confidence intervals, while the green shaded areas depict 90 percent confidence intervals. Standard errors are clustered at the local administrative unit level.

if physicians adopt effective treatment methods from their peers in tertiary hospitals, we would expect to observe better-managed care—such as reduced unnecessary testing and shorter hospital stays—while maintaining or improving patients' health conditions. This could lead to both clinical and economic gains, benefiting both patients and the healthcare system. On the other hand, if physicians indiscriminately adopt advanced but potentially unnecessary drugs or technologies, this could increase medical expenditures, such as higher testing frequency

Table 5
DID estimates of the policy's effect on inpatient duration.

	Dependent variable: Inpatient duration (days)			
	(1)	(2)	(3)	(4)
Post × Treat	−0.6903**	−0.7407**	−0.9807***	−0.9924***
	(0.2830)	(0.3026)	(0.3375)	(0.3418)
Pre-Mean	12.565	12.565	11.974	11.974
Year–Month FE	✓	✓	✓	✓
Hospital FE	✓		✓	
Disease FE	✓		✓	
Hospital–Disease FE		✓		✓
Individual controls	✓	✓	✓	✓
COVID-19 control	✓	✓	✓	✓
Radiological report required	✓	✓		
No. of Obs.	34,145	34,145	83,220	83,220
R^2	0.14	0.14	0.13	0.13

Note: This table reports the intent-to-treat estimates of the policy's effect on inpatient duration in secondary hospitals, focusing on the coefficient δ from the difference-in-differences specification in Eq. (2). The dependent variable measures inpatient duration (in days) among patients in secondary hospitals. Columns (1)–(2) report results for the subsample of six secondary hospitals that provided radiological reports, while columns (3)–(4) extend the analysis to all 13 secondary hospitals, including the seven hospitals that did not provide radiological reports. The corresponding event-study estimates are plotted in Subfigure (a) of Figure A7 in the Appendix. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

or prolonged hospitalization costs, ultimately imposing a greater financial burden on patients. The net effect of the UMA policy on patient outcomes, therefore, remains ambiguous and requires careful empirical investigation.

5.2.1. Inpatient duration

Ideally, evaluating the policy's effect on patient recovery would require long-term clinical follow-ups or post-discharge health measures. Such data are not available in our setting. We therefore follow the literature and use inpatient duration, measured by length of stay, as a proxy for recovery-related clinical efficiency. Under this framework, and conditional on patient characteristics and diagnosis, we interpret shorter inpatient stays as reflecting faster recovery and more efficient inpatient management, while longer stays indicate slower recovery.

Table 5 presents the results of this analysis. Columns (1)–(2) report estimates for our main sample, which includes the six secondary hospitals that provided radiology reports. Columns (3)–(4) extend the analysis to all 13 secondary hospitals in our dataset, encompassing both the six hospitals with radiology reports and the seven without. Across all specifications, the UMA policy is associated with a statistically significant reduction in inpatient stays. In our preferred specification, which adds hospital–disease and year–month fixed effects, the patient-level controls, and the city-month COVID-19 control (column 2), the average length of stay falls by 0.74 days against a pre-policy mean of 12.6 days—a reduction of roughly 6 percent. This effect is slightly larger in the full 13-hospital sample, where stays fall by 0.99 days from a pre-policy mean of 12.0 days (column 4), a decline of approximately 8 percent. These results indicate that patients in secondary hospitals were discharged sooner after the policy, consistent with faster recovery and more efficient inpatient management as physicians aligned their practice with tertiary-hospital protocols.[27]

[27] While we also examined 31-day readmission rates as an alternative proxy for patient health, this outcome is an extremely rare event in our data—occurring in only 2.9 percent of admissions in the full secondary sample and 0.1 percent in the primary radiology sample. Because the variable lacks sufficient variation to support reliable econometric inference, we exclude it from the main analysis, noting that the estimated policy effect is statistically indistinguishable from zero across all specifications.

Table 6
DID estimates of the policy's effect on imaging test utilization.

	Frequency of imaging tests		Likelihood of imaging test	
	(1)	(2)	(3)	(4)
Post × Treat	−0.0284	−0.0293	0.0077	0.0075
	(0.0699)	(0.0708)	(0.0052)	(0.0052)
Pre-Mean	0.678	0.678	0.914	0.914
Year–Month FE	✓	✓	✓	✓
Hospital–Disease FE	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓
COVID-19 control		✓		✓
No. of Obs.	83,220	83,220	83,220	83,220
R^2	0.40	0.40	0.50	0.50

Note: This table reports the intent-to-treat estimates of the policy's effect on imaging test utilization in secondary hospitals, focusing on the coefficient δ from the difference-in-differences specification in Eq. (2). Columns (1)–(2) report the effect on the number of imaging tests per patient, while Columns (3)–(4) report the effect on the likelihood of receiving any imaging test. The corresponding event-study estimates for the two margins are plotted in Subfigures (b) and (c) of Figure A7 in the Appendix. Standard errors are clustered at the hospital–disease level and reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

5.2.2. Imaging tests

We next investigate whether the UMA policy influenced the frequency of imaging tests administered to patients in secondary hospitals. A key concern is whether the observed improvements in prescription alignment between secondary and tertiary hospitals—highlighted in our main results—are driven by an increase in imaging tests at secondary hospitals. Such an increase could lead to over-testing, potentially reducing patient utility despite the policy's positive effects on treatment alignment. Conversely, if the improvements in prescription alignment were achieved without a rise in imaging test frequency, this would suggest that patient care at secondary hospitals has improved without imposing additional diagnostic burdens.

To address this question, we examine the test behavior along two margins. The intensive margin counts the number of imaging tests a patient receives, and the extensive margin records whether the patient receives any imaging test at all. We estimate both margins on all 13 secondary hospitals, drawing on the structured admission records that report imaging counts and incidence for every patient. The results, presented in Table 6, reveal no evidence of an increase in imaging test frequency either on the intensive margin or the extensive margin following the implementation of the UMA policy. These findings suggest that the observed improvements in treatment alignment were not accompanied by excessive imaging tests, alleviating concerns about potential over-testing and its associated negative effects on patients. Taken together, these results provide suggestive evidence that the UMA policy did not impose additional diagnostic burdens on patients through unnecessary imaging procedures.

5.2.3. Medical spending

Lastly, we analyze whether the UMA policy affected inpatient expenditures—a crucial determinant of both healthcare efficiency and patient welfare. While Section 5 demonstrated improved prescription alignment between secondary and tertiary hospitals, an important unanswered question remains: did this convergence lead to more cost-effective care or, conversely, to unnecessary increases in medical spending? Excessive spending—particularly if driven by unnecessary treatments—could diminish the overall benefits of the policy by increasing the financial burden on patients. To assess this, we examine three expenditure outcomes: total costs, medication expenses, and non-medication expenses (e.g., procedures, tests). As shown in Table 7, total costs fell significantly, by roughly a third. The decline lands almost entirely on non-medication spending, which dropped by about 30 percent, while medication spending shows no significant change.

Table 7
DID estimates of the policy's effect on medical spending.

	Total expenditure		Drug expenditure		Non-Drug expenditure	
	(1)	(2)	(3)	(4)	(5)	(6)
Post × Treat	−0.4291***	−0.4338***	−0.0939	−0.0926	−0.3601***	−0.3602***
	(0.1386)	(0.1393)	(0.0880)	(0.0902)	(0.0788)	(0.0798)
Year–Month FE	✓	✓	✓	✓	✓	✓
Hospital–Disease FE	✓	✓	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓	✓	✓
COVID-19 control		✓		✓		✓
No. of Obs.	79,734	79,734	69,328	69,328	68,763	68,763
R^2	0.21	0.21	0.23	0.23	0.19	0.19

Note: This table reports the intent-to-treat estimates of the policy's effect on medical spending in secondary hospitals, focusing on the coefficient δ from the difference-in-differences specification in Eq. (2). The dependent variables are expressed in logarithmic form as $\log(1+y)$, where y corresponds to the respective spending measures indicated in the column headers. Columns (1)–(2) report results for total medical spending, columns (3)–(4) examine total medication spending, and columns (5)–(6) focus on non-medication spending. The analysis is conducted across all 13 secondary hospitals in the sample, including six hospitals that provided radiological reports and seven that did not. The corresponding event-study estimates for the three spending measures are plotted in Subfigures (d)–(f) of Figure A7 in the Appendix. Standard errors are clustered at the hospital–disease level and reported in parentheses.*** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

These findings, combined with the evidence from Section 5.1 that secondary hospitals reduced the prescription of medications not aligned with benchmark standards, suggest that the UMA policy improved the precision of medication use through knowledge spillovers from tertiary hospitals. This not only enhances treatment quality but also alleviates the financial burden on patients, thereby improving overall healthcare efficiency and patient welfare.

6. Heterogeneity analysis

Our main results establish a core trade-off: the UMA policy successfully promoted practice alignment (exploitation) but likely at the cost of reduced academic output (exploration). Our next inquiry is how the two margins are linked. Under a simple workload channel, the additional clinical duties created by the policy would leave less time for research and produce a broad decline in academic output. By contrast, a protocol-dependence channel implies a more targeted exploitation–exploration trade-off. If external mentorship imposed a dominant and validated clinical protocol on physicians in secondary hospitals, the decline in independent inquiry should be concentrated in the domains most directly shaped by that protocol. We proceed in three steps. We first ask which physicians drove clinical convergence. We then split the publication decline by disease type. Finally, we run two tests of protocol dependence that trace the decline to individual physicians.

6.1. Practice alignment by physician volume

We first examine which physicians were most involved in practice adoption. The incentive to adopt the benchmark's practices should be strongest for high-volume physicians. These physicians account for a larger share of patient care, so standardized protocols are especially valuable to them as tools for improving clinical efficiency and reducing decision costs. The results, presented in Table 8, support this prediction. UMA increases practice similarity for both high- and low-volume physicians, but the response is substantially larger among high-volume physicians. For high-volume physicians, the estimated effect is approximately 0.020, compared with 0.012 for low-volume physicians. These patterns suggest that practice alignment was more pronounced among physicians with higher pre-policy patient volumes.

Table 8
Heterogeneous DID estimates of the policy's impact on practice adoption by physician experience.

	High-Volume Physicians		Low-Volume Physicians	
	(1)	(2)	(3)	(4)
Post × Treat	0.0196***	0.0198***	0.0120**	0.0120**
	(0.0020)	(0.0020)	(0.0053)	(0.0053)
Year–Month FE	✓	✓	✓	✓
Hosp-Disease FE	✓	✓	✓	✓
Individual controls	✓	✓	✓	✓
COVID-19 control		✓		✓
No. of Obs.	17 732	17 732	16 411	16 411
R^2	0.07	0.07	0.18	0.18

Note: This table presents heterogeneous DID estimates of the UMA policy's impact on prescription similarity scores, based on the specification in Eq. (2). The sample is split into two groups based on physician experience, proxied by their pre-policy patient volume (before June 2019). Columns (1)–(2) report results for "High-Volume Physicians," defined as those with a pre-policy patient volume above the sample median. Columns (3)–(4) report results for "Low-Volume Physicians," defined as those with a volume at or below the median. Even-numbered columns additionally control for the impact of the COVID-19 pandemic. Standard errors, reported in parentheses, are clustered at the hospital–disease level. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

6.2. Publication decline by disease type

Having identified the physicians most exposed to, and responsive to, the clinical standardization induced by UMA, we next examine where the decline in exploration appears. The chronic-disease margin is especially useful for separating the two explanations because China's hierarchical-care agenda placed routine chronic disease management increasingly on secondary hospitals. If the publication decline mainly reflected heavier clinical burdens, we would expect the decline to be most pronounced in chronic disease research, since secondary hospitals were increasingly expected to absorb routine chronic disease management. A protocol-dependence account instead predicts a larger decline in domains where external practices are more readily standardized and replicated, making independent exploration more vulnerable to convergence on a dominant external protocol (Haynes et al., 2009; Greenhalgh et al., 2017).

The estimates are not aligned with the workload prediction. If anything, the decline is concentrated outside chronic disease research. As shown in Table 9, UMA reduces non-chronic disease research across both quantity and quality margins: total publications fall by 7.7%, first-authored papers by 7.5%, high-impact journal publications by 4.1%, and citations by 15.1%. Similarly, Table A2 shows a similar pattern on the extensive margin, where the probability of producing non-chronic disease research declines significantly across the corresponding outcomes. By contrast, the estimates for chronic disease-related research are smaller and less precisely estimated. Figure A8 shows that this pattern is not driven by differential pre-trends; the non-chronic coefficients are flat before UMA and turn increasingly negative afterward. The decline therefore does not appear where a chronic-care workload explanation would most naturally predict it. Instead, it is concentrated in domains where clinical practice is more readily standardized and where convergence on external protocols may reduce the scope for independent exploration.

6.3. Two tests of protocol dependence

The disease-type split is difficult to reconcile with a general workload channel but, on its own, does not isolate protocol dependence as the relevant channel. Chronic and non-chronic research may come from different physicians and departments, and the two fields may differ in publication outlets, research cycles, and exposure to clinical disruptions. The larger non-chronic decline could therefore reflect differences between fields rather than a change in how individual physicians work. We therefore run two further tests that trace the

Table 9
Heterogeneous DID estimates of the policy's impact on academic outcomes by disease type.

	Chronic disease				Non-Chronic disease			
	Total (1)	First-authored (2)	Top journal (3)	Citations (4)	Total (5)	First-authored (6)	Top journal (7)	Citations (8)
Post × Treat	−0.0249*	−0.0308**	−0.0241***	−0.0590*	−0.0768***	−0.0752***	−0.0408***	−0.1506***
	(0.0128)	(0.0133)	(0.0081)	(0.0345)	(0.0219)	(0.0217)	(0.0078)	(0.0484)
Hospital FE	✓	✓	✓	✓	✓	✓	✓	✓
Year–Month FE	✓	✓	✓	✓	✓	✓	✓	✓
City linear trends	✓	✓	✓	✓	✓	✓	✓	✓
COVID-19 control	✓	✓	✓	✓	✓	✓	✓	✓
No. of Obs.	49 464	49 464	49 464	49 464	49 464	49 464	49 464	49 464
R^2	0.45	0.45	0.19	0.36	0.49	0.49	0.25	0.40

Note: This table presents the heterogeneous effects of the UMA policy implementation on physicians' academic outcomes in secondary hospitals, focusing on the coefficient γ from Eq. (4). Columns (1)–(4) capture publications related to chronic diseases, while columns (5)–(8) capture publications related to non-chronic disease topics. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

Table 10
Within-Author own-field versus Other-Field publication effects.

	Publication quantity		Publication quality	
	Total (1)	First-authored (2)	Top journal (3)	Citations (4)
Panel A. Publications in the physician's own-field set				
Post × Treat	−0.126***	−0.100***	−0.029***	−0.388***
	(0.041)	(0.032)	(0.010)	(0.091)
Pre-policy mean	1.135	0.622	0.020	5.061
Panel B. Publications outside the physician's own-field set				
Post × Treat	0.009	−0.009	0.000	−0.060
	(0.025)	(0.014)	(0.004)	(0.062)
Pre-policy mean	0.077	0.026	0.003	0.407
Author FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
City linear trends	✓	✓	✓	✓
No. of Obs. (author-year)	7520	7520	7520	7520

Note: Each panel reports a separate regression on the same author-year panel and differs only in the outcome variable. In Panel A, the outcome is the author–year number of publications classified within the physician's own-field set; in Panel B, it is the corresponding number of publications outside that set. The panel covers 2015–2022 and is restricted to physicians at secondary hospitals who published at least four papers during the pre-policy period from January 2015 through May 2019. The own-field set is the union of the physician's two most-published pre-policy departments (primary plus secondary specialty); other-field is all remaining departments. As an auxiliary check on the stability of the cross-field composition of research, we estimate a local-administrative-unit–month DID for the publication share of each of the twelve most common pre-policy departments. Together, these departments account for approximately 41% of pre-UMA publications. Eleven of the twelve estimated coefficients are statistically indistinguishable from zero. The only marginally significant coefficient, for Anesthesiology (0.020*), is positive and therefore does not support a compositional shift that could mechanically account for the decline in physicians' own-field publications. Standard errors, clustered at the local administrative unit level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

decline to individual physicians. The first test compares, within each physician, publications in the physician's own fields with publications outside them. The second test compares, across physicians in the same hospitals, those whose specialty coincides with the local flagship hospital's clinical strength with those whose specialty lies elsewhere. The two tests share one prediction. If UMA simply left physicians less time for research, output should fall broadly, across fields and across physicians. If transferred protocols crowd out independent inquiry, the decline should concentrate where those protocols apply most directly.

The own-field test. We define a physician's own-field set as the two departments in which the physician published most often before the policy and classify every publication as inside or outside that set. Within physicians, the decline is confined to their own fields. Publications inside the own-field set fall on all four outcomes, and every estimate in Panel A of Table 10 is statistically significant. Total own-field publications fall by 0.13 papers per physician per year, about one-tenth of the pre-policy mean. Publications outside the own-field set barely move, and none of the Panel B estimates is distinguishable from zero. A uniform loss of research time would not produce this asymmetry. City-wide shifts in the composition of research fields do not account for the pattern either (see the notes to Table 10 for the auxiliary test).

The specialty-match test. The own-field decomposition brings the analysis closer to physicians' clinical practice, but a workload-based explanation remains possible if UMA-induced clinical or training demands were concentrated in physicians' primary fields. We therefore further narrow the comparison by asking whether the decline is more pronounced when a physician's primary field overlaps with the clinical strength of the city's leading tertiary hospital. For each city, we designate the tertiary hospital with the largest pre-policy publication volume as the flagship hospital.[28] We classify a secondary-hospital physician as specialty matched when the physician's primary pre-policy specialty coincides with the flagship hospital's most-published department. Specialty-matched physicians are therefore those whose primary fields overlap with the domains in which locally available tertiary expertise is likely to be strongest.

The decline is larger where the match is closer, as Table 11 shows. Among specialty-matched physicians, total publications fall by 0.39 papers per physician per year, roughly three times the 0.13 decline among unmatched physicians in the same secondary hospitals. Citations fall by 1.7 per matched physician against 0.26 for unmatched. A uniform time squeeze does not explain this gradient. Additional clinical responsibilities may depress research overall, but they give no reason for the loss to concentrate in the one field the flagship hospital mentors most intensively. A within-hospital reallocation account is also difficult to reconcile with the results. If hospitals had shifted research effort away from mentored fields, unmatched physicians should have gained, and Panel B shows no such gains. The evidence instead points to a localized crowd-out of research in the domains where external clinical expertise is most relevant.

Taken together, the heterogeneity results reveal a consistent pattern. Clinical practice alignment is more pronounced among high-volume physicians, while the decline in academic output is concentrated in research on non-chronic conditions, physicians' primary fields, and specialties aligned with the leading tertiary hospital's clinical

[28] We manually verified the official institutional grade of each identified flagship hospital. In every sample city, the highest-publishing institution is officially designated as a Grade-A tertiary hospital, the highest tier in China's hospital accreditation system, consistent with the UMA policy's design of using regional elite providers as alliance leaders.

Table 11
Heterogeneity by specialty match to the flagship hospital.

	Publication quantity		Publication quality	
	Total (1)	First-authored (2)	Top journal (3)	Citations (4)
Panel A. Specialty-matched authors (N = 328 author-years)				
Post × Treat	−0.386***	−0.171**	−0.037	−1.739***
	(0.100)	(0.071)	(0.028)	(0.242)
Panel B. Specialty-unmatched authors (N = 5880 author-years)				
Post × Treat	−0.126*	−0.085*	−0.028*	−0.264*
	(0.071)	(0.048)	(0.016)	(0.136)
Author FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
City linear trends	✓	✓	✓	✓

Note: Panel A and Panel B report regressions on two mutually exclusive author subsamples, and the number of observations therefore differs across panels. The author–year panel covers 2015–2022 and is restricted to physicians at secondary hospitals who published at least four papers during the pre-policy period from January 2015 through May 2019. A physician's primary pre-policy specialty is defined as the clinical department in which the physician published most frequently during that period. For each city, the flagship hospital is the tertiary hospital in the same prefecture with the largest pre-policy CNKI publication volume, and its top department is the department in which it published the most pre-policy papers. Panel A includes specialty-matched physicians, whose primary pre-policy specialty equals the flagship hospital's top department. Panel B includes unmatched physicians, whose primary pre-policy specialty lies outside the flagship hospital's five most-published departments. Physicians whose primary pre-policy specialty ranks second through fifth in the flagship hospital's department list are excluded from both panels. Standard errors, clustered at the local administrative unit level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.1$.

strength. This pattern is difficult to reconcile with a simple time-squeeze explanation, which would predict a broader decline across physicians and research fields. Instead, the estimated publication decline becomes more pronounced as the analysis focuses on domains in which UMA mentorship, training, and protocol transfer were likely to be most relevant. The evidence therefore suggests that the same external standards that helped secondary hospitals converge toward the benchmark also constrained the scope for independent research in the most directly mentored areas. Over time, this form of practice standardization may reduce idiosyncratic clinical variation, but it may also make secondary-hospital research more dependent on the priorities and protocols of leading tertiary mentors.

7. Robustness checks

Our difference-in-differences design compares changes in outcomes in UMA pilot areas before and after policy implementation with contemporaneous changes in non-pilot areas. A central identification concern is whether factors other than UMA could have generated the observed divergence between treated and control areas. Potential threats include differential pre-existing trajectories, concurrent health-sector reforms, pandemic-related disruptions, and sensitivity to alternative modeling choices. To systematically address these concerns and assess the credibility of our empirical design, we conduct a battery of falsification and robustness checks. Our main findings remain robust across these exercises.

Parallel trends and trend sensitivity. The event-study estimates provide no evidence of differential pre-trends. In both the practice-alignment event study (Fig. 3) and the annual publication event study (Fig. 4), the pre-policy coefficients are statistically indistinguishable from zero, whereas the post-policy coefficients begin to diverge from zero only after UMA implementation. We further assess whether the results are sensitive to alternative assumptions about the underlying trajectories of treatment and control units.

For practice alignment, our preferred specification intentionally omits hospital-specific linear trends to avoid improperly absorbing dynamic treatment responses. The results are nevertheless robust to introducing either hospital-specific or city-specific linear trends, which yields only the modest attenuation expected from over-fitting (Table A3). We also follow Dustmann et al. (2022) by estimating the differential pre-policy trend and removing its projected continuation from the post-policy outcome. This adjustment yields estimates close to the baseline results (Table A4), indicating that the estimated effect is not driven by the continuation of differential pre-policy trends. Finally, we allow cities with different pre-policy patient compositions along dimensions of age, gender, marital status, and insurance coverage to follow flexible differential trajectories by interacting these baseline characteristics with year–month fixed effects (Table A5). Across these demanding specifications the practice-alignment estimate remains robust and strongly significant.

We apply a similar set of trend sensitivity tests to the publication outcomes. Unlike high-frequency clinical decisions, academic research develops over longer horizons and may be influenced by slow-moving changes in regional R&D investment and institutional capacity. For this reason, our baseline publication model already includes city-specific linear trends to absorb these gradual shifts in the local research environment. To allow underlying research trajectories to vary flexibly with baseline city capacity, we interact five city characteristics measured in 2018—population, GDP per capita, and the numbers of hospitals, beds, and physicians—with quadratic and cubic functions of calendar time (Table A6). The estimated effects remain close to the baseline results. Across these specifications, total publications fall by 6 to 7 percent and citations by 11 to 13 percent, compared with baseline estimates of 7 and 13 percent, respectively. Together, these results indicate that the observed decline in research output is not driven by pre-existing regional trajectories.

Placebo and falsification. A natural falsification test emerges from the targeted structure of the UMA policy. Specifically, while secondary hospitals are the designated recipients of UMA mentorship, leading tertiary hospitals act primarily as the suppliers. This asymmetric treatment assignment provides an ideal in-sample placebo. If the baseline results were driven by broad, city-wide macroeconomic shocks or concurrent healthcare reforms rather than the UMA intervention itself, tertiary hospitals in pilot cities should exhibit similar behavioral changes. We find no evidence of such confounding shocks. For clinical practice, tertiary physicians show no convergence in prescribing under any specification (Table A7), distinct from the patterns documented for the secondary tier.

We find similarly null effects for research output. Re-estimating the baseline difference-in-differences model for tertiary hospitals yields small and statistically insignificant estimates across all four publication outcomes (Table A8). The corresponding event-study coefficients also remain close to zero throughout the post-policy period (Figure A9). These findings are informative because administrative pressure and incentives for academic publishing are more strongly institutionalized at tertiary hospitals. A city-level initiative affecting research activity across hospital tiers would therefore also be expected to leave a detectable effect among tertiary hospitals. The null results weaken the concern that the decline observed among secondary hospitals was driven by tier-neutral city-level policies or shocks.

Concurrent policy shocks. A central concern is whether contemporaneous national reforms, rather than UMA, account for the observed changes in clinical practice and academic research. Because the clinical-practice and academic-exploration outcomes are potentially affected by different concurrent reforms, we consider the relevant policy shocks for each outcome in turn.

For the clinical-practice outcome, the most relevant concurrent reform is the National Drug Price Negotiation (NDPN). The negotiation round relevant to our sample took effect in January 2020 and

substantially reduced the prices of covered drugs.[29] The reform could affect our prescription-based alignment measure if hospitals responded by changing the therapeutic categories from which they prescribed. Our similarity score, however, is constructed from a 32-dimensional vector of broad therapeutic categories and is therefore unaffected by substitutions among specific products within the same category. A remaining concern is that hospitals in pilot areas may have responded differentially to the NDPN by shifting prescriptions across therapeutic categories. To assess this possibility, we jointly estimate the baseline model with a post-NDPN indicator interacted with treatment. The UMA coefficient remains similar in magnitude and statistically significant, while the NDPN interaction is statistically indistinguishable from zero (Table A9). These results provide no evidence that differential responses to the NDPN account for the estimated increase in practice alignment.

For the academic exploration outcomes, the relevant threats are two national reforms that explicitly altered research incentives: the February 2020 cancellation of cash bonuses for SCI-indexed papers and the July 2021 Health Professional Title Reform (HPTR).[30] While their aggregate effects are naturally absorbed by the year–month fixed effects, differential local enforcement remains a concern. We address this concern in two ways. First, Panel A of Tables A10 and A11 jointly estimates the UMA effect and the relevant post-reform indicator interacted with treatment. Among secondary hospitals, the UMA estimate remains negative across all four publication outcomes and statistically significant on most margins, while the SCI and HPTR interactions show no systematic pattern across outcomes. Second, Panel B of the same tables reports the corresponding across-tier placebo. Under the maintained assumption that these national reforms did not target hospital tiers differently, stronger implementation in pilot areas should generate a detectable publication response at both secondary and tertiary hospitals. We find no such response at the tertiary tier: both the UMA and reform interactions are close to zero and statistically insignificant across publication outcomes. The joint estimates and across-tier placebo weaken the concern that the decline among secondary hospitals is driven by differential implementation of contemporaneous publication-incentive reforms rather than by UMA.

The COVID-19 pandemic. Another important concern is that the onset of the COVID-19 pandemic overlapped with the post-policy period. While our baseline specifications directly control for city-month COVID-19 case counts, reported infections may not fully capture broader disruptions to healthcare delivery and research activity, including shifts toward remote consultation, the reallocation of medical resources, and reductions in patient flows. Because the clinical-practice and publication outcomes are measured at different frequencies and may respond to the pandemic through different channels, we assess this concern separately for the two domains.

For the high-frequency practice-alignment outcome, we can directly exclude the pandemic period by restricting the sample to observations before January 2020. The estimate remains very close to the full-sample baseline: the pre-COVID estimate is 0.0177, compared with the baseline estimate of 0.0202 (Table A12). The event-study estimates provide complementary evidence on timing (Fig. 3). Practice alignment begins to increase during the post-UMA but pre-COVID period, while the pandemic period is associated mainly with greater volatility rather than the initial emergence of the effect. These results show that the increase in practice alignment was already present before the onset of COVID-19 and therefore cannot be attributed entirely to pandemic-era changes in clinical practice.

For the publication outcomes, an analogous pre-COVID restriction is less informative because research output is realized with substantial production and editorial lags, and dropping the post-2020 period would remove much of the relevant post-treatment variation. We therefore rely on several complementary exercises. First, following the classification of Fang et al. (2020), all our sample cities fall within the mildest pandemic-control category, reducing the concern that the estimates compare severely restricted cities with relatively unaffected ones. In this setting, our baseline specification addresses pandemic exposure through two layers: year–month fixed effects absorb national shocks common to all cities, and city-month COVID-19 case counts capture differences in local infection intensity. Second, to control for broader behavioral disruptions independent of case counts, we include a Baidu city-month mobility index that captures realized population movement. The UMA estimates remain similar, and the mobility coefficient is statistically indistinguishable from zero (Table A13). Third, the estimated decline in publication output remains when we exclude 2020, and the timing of the effects does not indicate that they emerge only during the peak pandemic period (Table A14). Finally, the across-tier placebo provides an additional comparison. Tertiary hospitals located in the same areas were exposed to the same city-level pandemic conditions but exhibit no corresponding decline in research output (Table A8 and Figure A9). Taken together, these exercises weaken the concern that the decline in secondary-hospital publications is driven by differential exposure to the COVID-19 pandemic.

Alternative approaches for zero-inflated outcomes. Physician publication counts contain a high proportion of zeros (Figure A6). Because our main analysis models these outcomes using a $\log(y+1)$ transformation, recent literature raises concerns that this approach may distort estimates along the extensive margin (Chen and Roth, 2024). To address these concerns and validate our results, we employ two alternative estimation strategies tailored to the nature of our data.

First, we estimate a Poisson regression model for count outcomes, as advocated by Chen and Roth (2024). The results, reported in Table A15, show statistically significant reductions across all four measures of academic productivity. The coefficient estimates imply declines of 15.4% in total publications, 15.6% in first-authored publications, 32.6% in top-tier journal publications, and 16.8% in citations.[31] These estimates are consistent in direction with our primary results and confirm that the estimated decline in physician research output is not driven by the $\log(y+1)$ transformation.

Second, we estimate linear regressions using the untransformed levels of physician publications. This approach avoids the potential distortions introduced by log transformations and provides a straightforward interpretation of the policy's impact. As reported in Table A16, the UMA policy continues to exert a significant negative effect on academic output in secondary hospitals. Expressed as average proportional treatment effects (Chen and Roth, 2024), we estimate that the UMA

[29] The NDPN is conducted in recurring rounds by the National Healthcare Security Administration, with negotiated prices incorporated into the National Reimbursement Drug List and generally taking effect in the following year. The round relevant to our practice-alignment sample was announced in November 2019 and took effect on January 1, 2020. We therefore date the shock to January 2020. Earlier rounds took effect before UMA implementation, while subsequent rounds fall outside our practice-alignment window.

[30] The February 2020 SCI-related reform directed institutions to stop using SCI-related indicators as a basis for monetary rewards, reducing the financial return to publication (http://www.moe.gov.cn/srcsite/A16/moe_784/202002/t20200223_423334.html). The July 2021 Health Professional Title Reform de-emphasized publications in physician title and promotion review by allowing clinical case records and other representative works to count in evaluation, reducing the career return to publication-related research (https://www.gov.cn/zhengce/zhengceku/2021-08/05/content_5629566.htm). Both reforms are therefore natural candidate confounders for the exploration decline.

[31] For the Poisson specification, the proportional treatment effect is calculated as $\exp(\hat{\beta})-1$. Because the estimated coefficients are negative, we report the magnitudes of the implied declines as $1-\exp(\hat{\beta})$. The calculations are $1-\exp(-0.1675)=15.4\%$ for total publications, $1-\exp(-0.1699)=15.6\%$ for first-authored publications, $1-\exp(-0.3941)=32.6\%$ for top-tier journal publications, and $1-\exp(-0.1845)=16.8\%$ for citations.

policy reduces total publications by 28.2%, first-authored publications by 30.1%, top-tier journal publications by 49.4%, and citations by 40.6%.[32] These proportional estimates further substantiate the negative impact of the policy on physician productivity and align closely with the results from the Poisson regression model.

Overall, the main estimates are robust to alternative specifications of underlying trends, across-tier placebo tests, explicit controls for concurrent reforms and pandemic-related disruptions, and alternative estimators for zero-inflated publication outcomes. The absence of corresponding effects among tertiary hospitals further reduces the plausibility of tier-neutral city-level policies or shocks as an explanation for the secondary-hospital results. Taken together, the evidence supports the interpretation that UMA increased practice alignment while reducing academic output in secondary hospitals, rather than these patterns reflecting pre-existing trajectories, overlapping shocks, or particular modeling choices.

8. Conclusion

This study confronts a fundamental tension in knowledge-intensive sectors, particularly in healthcare: does promoting the exploitation of existing best practices help or hinder the exploration of new ideas? We investigate this question in the context of China's Urban Medical Alliance (UMA) policy, which was designed to facilitate knowledge spillovers from elite tertiary hospitals to secondary hospitals. Leveraging the designation and timing of UMA pilot areas in a difference-in-differences framework, we combine conventional administrative outcomes with a language-model-based measure of clinical practice alignment constructed from radiological reports and prescription records. We find that UMA increased the alignment of prescribing practices in secondary hospitals and improved inpatient efficiency. At the same time, physicians in treated hospitals experienced declines in academic output, particularly in top-tier journals and in the citation impact of their work, indicating a suppression of innovative activity. These findings document a trade-off between the diffusion of established clinical knowledge and the production of new academic knowledge.

Our findings make three contributions. First, this study extends the study of knowledge spillovers to the healthcare sector and highlight potentially asymmetric effects on recipient institutions. Unlike industries where spillovers typically enhance both efficiency and innovation, our findings highlight the unique challenges in healthcare, where the adoption of standardized protocols can limit opportunities for creative exploration. Second, the study provides evidence that the diffusion of established knowledge can reshape the balance between exploitation and exploration within recipient organizations. The results are consistent with the idea that greater reliance on externally validated practices may reduce the incentives or scope for independent inquiry. This insight is particularly salient for knowledge-intensive sectors, where balancing efficiency and creativity is essential for sustained progress. Third, we develop a scalable approach to measuring clinical practice alignment from unstructured medical text. By using contrastively trained language-model embeddings to identify clinically comparable cases and then comparing their prescription patterns, the method complements measures based on coarse administrative codes and may be useful for studying the diffusion of complex professional knowledge in other settings.

The policy implications should be interpreted in light of both sides of this trade-off. Our results do not imply that programs promoting best-practice diffusion are undesirable. UMA narrowed differences in clinical practice and improved dimensions of inpatient care. However, these gains were accompanied by lower research activity in recipient hospitals, particularly in fields most closely connected to the transferred expertise. Policymakers may therefore consider complementing clinical mentorship and protocol standardization with measures that preserve incentives for local inquiry. Possible approaches include protected research time, dedicated research funding for non-tertiary institutions, and evaluation systems that reward both the effective adoption of established practices and context-specific experimentation. Ultimately, sustainable progress in healthcare requires a delicate balance between perfecting what is known and discovering what is not. Determining which complementary policies can preserve exploration without weakening the gains from knowledge diffusion remains an important question for future research.

Several limitations also point to directions for further work. Our publication measures capture formal academic output but do not encompass all forms of clinical experimentation, informal learning, or innovation in care delivery. Further research could also incorporate more direct patient-level health outcomes, such as long-term mortality and recovery rates, to conduct a more comprehensive welfare analysis of the trade-off we identify. Finally, the tension between standardization and creativity is not unique to healthcare. Applying this framework to other knowledge-intensive sectors, such as education, law, or technology, could yield broader insights into managing the diffusion of knowledge in a way that fosters, rather than stifles, innovation.

CRediT authorship contribution statement

Yinan Liu: Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Chen Lin:** Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Conceptualization. **Zixiao Li:** Software, Resources, Data curation. **Jing Zhang:** Software, Resources, Data curation.

Funding

Zixiao Li: This research was supported by the Noncommunicable Chronic Diseases—National Science and Technology Major Project (Grant No. 2025ZD0546500 & 2025ZD0546504).

Yinan Liu: The author received no specific funding for this work.

Jing Zhang: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Chen Lin: This research was supported by the Major Program of the National Social Science Fund of China (Grant No. 22&ZD092).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jdeveco.2026.103892.

Data availability

The authors do not have permission to share data.

[32] Following Chen and Roth (2024), the proportional ATT in levels is calculated as $\theta_{ATT\%} = \frac{E[Y_{it}(1)|D_i=1,Post_t=1]-E[Y_{it}(0)|D_i=1,Post_t=1]}{E[Y_{it}(0)|D_i=1,Post_t=1]}$.

References

Abaluck, Jason, Agha, Leila, Kabrhel, Chris, Raja, Ali, Venkatesh, Arjun, 2016. The determinants of productivity in medical testing: Intensity and allocation of care. Am. Econ. Rev. 106 (12), 3730–3764.

Acemoglu, Daron, Akcigit, Ufuk, 2012. Intellectual property rights policy, competition and innovation. J. Eur. Econ. Assoc. 10 (1), 1–42.

Acemoglu, Daron, Gancia, Gino, Zilibotti, Fabrizio, 2012. Competing engines of growth: Innovation and standardization. J. Econom. Theory 147 (2), 570–601.

Adams, James D., 1990. Fundamental stocks of knowledge and productivity growth. J. Political Econ. 98 (4), 673–702.

Agha, Leila, Molitor, David, 2018. The local influence of pioneer investigators on technology adoption: evidence from new cancer drugs. Rev. Econ. Stat. 100 (1), 29–44.

Aghion, Philippe, Bloom, Nick, Blundell, Richard, Griffith, Rachel, Howitt, Peter, 2005. Competition and innovation: An inverted-U relationship. Q. J. Econ. 120 (2), 701–728.

Aghion, Philippe, Cai, Jing, Dewatripont, Mathias, Du, Luosha, Harrison, Ann, Legros, Patrick, 2015. Industrial policy and competition. Am. Econ. J.: Macroecon. 7 (4), 1–32.

Aghion, Philippe, Dewatripont, Mathias, Stein, Jeremy C., 2008. Academic freedom, private-sector focus, and the process of innovation. Rand J. Econ. 39 (3), 617–635.

Akcigit, Ufuk, Kerr, William R., 2018. Growth through heterogeneous innovations. J. Political Econ. 126 (4), 1374–1443.

Arthur, W. Brian, 1989. Competing technologies, increasing returns, and lock-in by historical events. Econ. J. 99 (394), 116–131.

Azoulay, Pierre, Graff Zivin, Joshua S., Manso, Gustavo, 2011. Incentives and creativity: evidence from the academic life sciences. Rand J. Econ. 42 (3), 527–554.

Banerjee, Abhijit, Duflo, Esther, Qian, Nancy, 2020. On the road: Access to transportation infrastructure and economic growth in China. J. Dev. Econ. 145, 102442.

Bloom, Nicholas, Eifert, Benn, Mahajan, Aprajit, McKenzie, David, Roberts, John, 2013. Does management matter? Evidence from India. Q. J. Econ. 128 (1), 1–51.

Bloom, Nicholas, Van Reenen, John, 2007. Measuring and explaining management practices across firms and countries. Q. J. Econ. 122 (4), 1351–1408.

Brynjolfsson, Erik, Li, Danielle, Raymond, Lindsey, 2025. Generative AI at work. Q. J. Econ. 140 (2), 889–942.

Cai, Jing, Szeidl, Adam, 2018. Interfirm relationships and business performance. Q. J. Econ. 133 (3), 1229–1282.

Chandra, Amitabh, Skinner, Jonathan, 2012. Technology growth and expenditure growth in health care. J. Econ. Lit. 50 (3), 645–680.

Chandra, Amitabh, Staiger, Douglas O., 2007. Productivity spillovers in health care: evidence from the treatment of heart attacks. J. Political Econ. 115 (1), 103–140.

Chandra, Amitabh, Staiger, Douglas O., 2020. Identifying sources of inefficiency in healthcare. Q. J. Econ. 135 (2), 785–843.

Chen, Yi, Fang, Hanming, Zhao, Yi, Zhao, Zibo, 2024. Recovering Overlooked Information in Categorical Variables with LLMs: An Application to Labor Market Mismatch. National Bureau of Economic Research.

Chen, Jiafeng, Roth, Jonathan, 2024. Logs with zeros? Some problems and solutions. Q. J. Econ. 139 (2), 891–936.

Cheng, Ruidong, Ye, Xiangming, Li, Juebao, Zhang, Jie, Zhang, Li, Fang, Shan, 2021. Survey on the research and teaching ability of rehabilitation based on public hospitals in Zhejiang province. Chin. J. Med. Sci. Res. Manag. (1), 57–62, In Chinese.

Clements, Matthew T., Ohashi, Hiroshi, 2005. Indirect network effects and the product cycle: video games in the US, 1994–2002. J. Ind. Econ. 53 (4), 515–542.

Currie, Janet, MacLeod, W. Bentley, 2017. Diagnosing expertise: Human capital, decision making, and performance among physicians. J. Labor Econ. 35 (1), 1–43.

Cutler, David, Skinner, Jonathan S, Stern, Ariel Dora, Wennberg, David, 2019. Physician beliefs and patient preferences: a new look at regional variation in health care spending. Am. Econ. J.: Econ. Policy 11 (1), 192–221.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186.

Dong, Shengjie, Millar, Ross, Shi, Chenshu, Dong, Minye, Xiao, Yuyin, Shen, Jie, Li, Guohong, 2021. Rating hospital performance in China: Review of publicly available measures and development of a ranking system. J. Med. Internet Res. 23 (6), e17095.

Dustmann, Christian, Lindner, Attila, Schönberg, Uta, Umkehrer, Matthias, vom Berge, Philipp, 2022. Reallocation effects of the minimum wage. Q. J. Econ. 137 (1), 267–328.

Fang, Hanming, Li, Ming, Lu, Guangli, 2025. Decoding China's Industrial Policies. National Bureau of Economic Research.

Fang, Hanming, Wang, Long, Yang, Yang, 2020. Human mobility restrictions and the spread of the novel coronavirus (2019-nCoV) in China. J. Public Econ. 191, 104272.

Finkelstein, Amy, Gentzkow, Matthew, Williams, Heidi, 2016. Sources of geographic variation in health care: Evidence from patient migration. Q. J. Econ. 131 (4), 1681–1726.

Furman, Jeffrey L., Stern, Scott, 2011. Climbing atop the shoulders of giants: The impact of institutions on cumulative research. Am. Econ. Rev. 101 (5), 1933–1963.

Gao, Tianyu, Yao, Xingcheng, Chen, Danqi, 2021. Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821.

Gentzkow, Matthew, Kelly, Bryan, Taddy, Matt, 2019. Text as data. J. Econ. Lit. 57 (3), 535–574.

Greenhalgh, Trisha, Wherton, Joseph, Papoutsi, Chrysanthi, Lynch, Jennifer, Hughes, Gemma, Hinder, Susan, Fahy, Nick, Procter, Rob, Shaw, Sara, et al., 2017. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J. Med. Internet Res. 19 (11), e367.

Halaburda, Hanna, Jullien, Bruno, Yehezkel, Yaron, 2020. Dynamic competition with network externalities: how history matters. Rand J. Econ. 51 (1), 3–31.

Han, Yi, Wu, Mingqin, 2024. Inter-regional barriers and economic growth: Evidence from China. J. Dev. Econ. 167, 103197.

Handel, Benjamin R., 2013. Adverse selection and inertia in health insurance markets: When nudging hurts. Am. Econ. Rev. 103 (7), 2643–2682.

Hao, Qian, Ji, Hao, Hu, Li, Yu, Yingying, 2023. Analysis on scientific research competitiveness of grade a second-class hospitals in Zhejiang province. J. Med. Inform. 44 (6), 59–64, In Chinese.

Haynes, Alex B, Weiser, Thomas G, Berry, William R, Lipsitz, Stuart R, Breizat, Abdel-Hadi S, Dellinger, E Patchen, Herbosa, Teodoro, Joseph, Sudhir, Kibatala, Pascience L, Lapitan, Marie Carmela M, et al., 2009. A surgical safety checklist to reduce morbidity and mortality in a global population. N. Engl. J. Med. 360 (5), 491–499.

Holmström, Bengt, 1999. Managerial incentive problems: A dynamic perspective. Rev. Econ. Stud. 66 (1), 169–182.

Hsieh, Chang-Tai, Klenow, Peter J., 2009. Misallocation and manufacturing TFP in China and India. Q. J. Econ. 124 (4), 1403–1448.

Hu, Hongwei, Liang, Hailun, Wang, Hufeng, 2021. Longitudinal study of the earliest pilot of tiered healthcare system reforms in China: will the new type of chronic disease management be effective? Soc. Sci. Med. 285, 114284.

Jacob, Hannah, 2016. Why all doctors should be involved in research. BMJ 352.

Jaffe, Adam B., Trajtenberg, Manuel, Henderson, Rebecca, 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.

Jaravel, Xavier, Petkova, Neviana, Bell, Alex, 2018. Team-specific capital and innovation. Am. Econ. Rev. 108 (4–5), 1034–1073.

Jin, Lawrence, Tang, Rui, Ye, Han, Yi, Junjian, Zhong, Songfa, 2024. Path dependency in physician decisions. Rev. Econ. Stud. 91 (5), 2916–2953.

Juhász, Réka, Lane, Nathan, Oehlsen, Emily, Pérez, Verónica, 2025. Measuring Industrial Policy: A Text-Based Approach. NBER Working Paper (w33895).

Kosmyna, Nataliya, Hauptmann, Eugene, Yuan, Ye Tong, Situ, Jessica, Liao, Xian-Hao, Beresnitzky, Ashly Vivian, Braunstein, Iris, Maes, Pattie, 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv preprint arXiv:2506.08872.

Levinthal, Daniel A., March, James G., 1993. The myopia of learning. Strateg. Manag. J. 14 (S2), 95–112.

Li, Xi, Krumholz, Harlan M, Yip, Winnie, Cheng, Kar Keung, De Maeseneer, Jan, Meng, Qingyue, Mossialos, Elias, Li, Chuang, Lu, Jiapeng, Su, Meng, et al., 2020. Quality of primary health care in China: challenges and recommendations. Lancet 395 (10239), 1802–1812.

Li, Xi, Lu, Jiapeng, Hu, Shuang, Cheng, KK, De Maeseneer, Jan, Meng, Qingyue, Mossialos, Elias, Xu, Dong Roman, Yip, Winnie, Zhang, Hongzhao, et al., 2017. The primary health-care system in China. Lancet 390 (10112), 2584–2594.

Li, Danielle, Raymond, Lindsey R., Bergman, Peter, 2026. Hiring as exploration. Rev. Econ. Stud. 93 (2), 1200–1240.

Li, Dan, Yu, Jing, Lv, Zhong-wei, Gu, Wen-Jun, Li, Ji-yu, 2022. Scientific research competitiveness in hospitals: a narrative review of major hospital ranking systems in China. Health Sci. Rep. 5 (3), e583.

Loughran, Tim, McDonald, Bill, 2016. Textual analysis in accounting and finance: A survey. J. Account. Res. 54 (4), 1187–1230.

Lu, Chen, Zhang, Zixiao, Lan, Xiuting, 2019. Impact of China's referral reform on the equity and spatial accessibility of healthcare resources: a case study of Beijing. Soc. Sci. Med. 235, 112386.

Manso, Gustavo, 2011. Motivating innovation. J. Finance 66 (5), 1823–1860.

March, James G., 1991. Exploration and exploitation in organizational learning. Organ. Sci. 2 (1), 71–87.

Ministry of Health of the People's Republic of China, 1989. Hierarchical management standards for general hospitals.

Molitor, David, 2018. The evolution of physician practice styles: evidence from cardiologist migration. Am. Econ. J.: Econ. Policy 10 (1), 326–356.

Mullahy, John, Norton, Edward C., 2022. Why Transform Y? A Critical Assessment of Dependent-Variable Transformations in Regression Models for Skewed and Sometimes-Zero Outcomes. National Bureau of Economic Research.

Murray, Fiona, Stern, Scott, 2007. Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. J. Econ. Behav. Organ. 63 (4), 648–687.

National Health Commission of the People's Republic of China, 2021. China health statistics yearbook.

Pan, Jay, Liu, Dan, Ali, Shehzad, 2015. Patient dissatisfaction in China: what matters. Soc. Sci. Med. 143, 145–153.
Pan, Jay, Zhao, Hanqing, Wang, Xiuli, Shi, Xun, 2016. Assessing spatial access to public and private hospitals in Sichuan, China: The influence of the private sector on the healthcare geography in China. Soc. Sci. Med. 170, 35–45.
Rambachan, Ashesh, Roth, Jonathan, 2023. A more credible approach to parallel trends. Rev. Econ. Stud. 90 (5), 2555–2591.
Reimers, Nils, Gurevych, Iryna, 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. EMNLP-IJCNLP, pp. 3982–3992.
Rochet, Jean-Charles, Tirole, Jean, 2003. Platform competition in two-sided markets. J. Eur. Econ. Assoc. 1 (4), 990–1029.
Rothaermel, Frank T., Deeds, David L., 2004. Exploration and exploitation alliances in biotechnology: A system of new product development. Strateg. Manag. J. 25 (3), 201–221.
Tsauo, Jiaywei, 2013. Employment pressure and the burden of publication in China. BMJ 347.
Vandenbroucke, Jan P., 1998. Medical journals and the shaping of medical knowledge. Lancet 352 (9145), 2001–2006.
Williams, Heidi L., 2013. Intellectual property rights and innovation: Evidence from the human genome. J. Political Econ. 121 (1), 1–27.
Ye, Jiangfeng, Jiang, Xue, Jing, Qi, Lei, Yi, 2019. International comparison of integrated healthcare service models and their implications. Manag. Rev. 31 (6), 199–212, In Chinese.
Yip, Winnie, Hsiao, William C., 2008. The Chinese health system at a crossroads. Health Aff. 27 (2), 460–468.
Yuan, Heng-Feng, Xu, Wei-Dong, Hu, Hai-Yan, 2013. Young Chinese doctors and the pressure of publication. Lancet 381 (9864), e4.
Zhang, Zhongheng, Winston, Gavin P., Zhao, Hai-Tao, Oei, Edwin H.G., Ai, Qiyun, Lofroy, Romaric, Lin, Tong, Shen, Yuxin, Ng, Chuen Ki, Liu, Huazhang, Civelek, A.C., Han, Zhijun, He, Yumin, Ji, Liying, W'ang, Yi-Xi'ang J., 2014. Focus on China: should clinicians engage in research? and lessons from other countries. Quant. Imaging Med. Surg. 4 (5), 413–425.
Zhou, Zhongliang, Zhao, Yaxin, Shen, Chi, Lai, Sha, Nawaz, Rashed, Gao, Jianmin, 2021. Evaluating the effect of hierarchical medical system on health seeking behavior: a difference-in-differences analysis in China. Soc. Sci. Med. 268, 113372.