Received: 7 September 2022 | Revised: 5 May 2023 | Accepted: 15 May 2023

DOI: 10.1111/gove.12804

Governance WILEY

ORIGINAL ARTICLE

Unpack the black box of pilot sampling in policy experimentation: A qualitative comparative analysis of China's public hospital reform

Alex Jingwei He[1] | **Yumeng Fan**[1] | **Rui Su**[2]

[1]Division of Public Policy, The Hong Kong University of Science and Technology, Hong Kong, China

[2]School of Government, Sun Yat-sen University, Guangzhou, China

Correspondence
Alex Jingwei He.
Email: ajwhe@ust.hk

Funding information
General Research Fund, Hong Kong Research Grants Council, Grant/Award Number: 18605720

Abstract

Governments increasingly use policy experimentation programs to seek solutions for complex problems. Because randomization and controllability are unrealistic for real-world policy experiments, how subnational pilots are selected is crucial for generating sound evidence for national replication. However, the received wisdom on pilot sampling is thin and paradoxical. While some studies suggest that policymakers prefer to select regions with favorable conditions, others contend that securing representativeness remains the principal concern when it comes to pilot selection. This study resolves the paradox by elucidating the logic of selecting pilots in large policy experimentation programs. We focus on China's huge public hospital reform program and through a novel research design that combines comparative qualitative analysis and illustrative case studies we seek to explain the strategy for pilot selection. Our analyses reveal five distinctive pathways of pilot sampling: *piloting for challenge*, *piloting for advancement*, *piloting for innovation*, *piloting for action*, and *piloting for regional generalization*. Each modality represents a specific experimental purpose. We reveal

that piloting serves as a versatile governance tool that can fulfill multiple functions in complex reforms.

1 | INTRODUCTION

Frequent and extensive use of policy experiments stands as a prominent feature of governance in contemporary societies. Such experiments help reduce uncertainties in policymaking by generating evidence and knowledge, and they promote learning and adaptive policy responses based on experience gained over time (Ansell & Bartenberger, 2016; Nair, 2020). Policy experiments also serve the purpose of demonstration, accelerating implementation and building acceptance in the community (Ettelt et al., 2015). In large countries with a multilevel governance structure, national policymakers increasingly resort to experimental programs across subnational regions, in their search for scalable solutions to complex policy problems (Ettelt et al., 2022; Husain, 2017; Jowell, 2003; Nair & Howlett, 2016). The recent literature has paid due attention to this thriving form of experimentalist governance, in which national and subnational governments interact in a dynamic fashion throughout the experimental process (He et al., 2022; Sabel & Zeitlin, 2010; Zhu & Zhao, 2021). Scholars have also gained an understanding, albeit preliminary, of the motives and behavioral patterns of subnational units when joining national-level experimental programs (Ettelt et al., 2022; Wang et al., 2022; Wang & Yang, 2021).

However, a blind spot in the literature is how and why subnational units are *selected* to participate in such programs. This is not a trivial matter of concern, because "sampling strategy" is of strategic importance in an experiment and fundamentally affects both internal validity and external validity, in common scientific language. We do know that randomization and controllability in lab settings are neither realistic nor suitable for real-world public policies (Ko & Shin, 2017; Nair, 2020; van der Heijden, 2014), but we know little regarding the criteria and considerations that national governments use when they sample subnational "subjects" in an experimental program. *Do they strive to secure "representativeness" by recruiting a sample that is as diverse as possible? Or, do they strategically oversample "subjects" with favorable conditions in order to maximize the chance of experimental success?* The evidence thus far is thin and inconclusive. Hence, this study sought to resolve the paradox using a case study of China's experimental program in healthcare reforms.

The Chinese government launched a national public hospital reform program in 2010, striving to overhaul the country's hospital system. In the absence of clear policy solutions, the government initiated an experimental program that eventually enlisted 200 cities nationwide (Li & Fu, 2017), thereby offering us an excellent opportunity to investigate the logic behind pilot sampling in a large governance system. Through comparative qualitative analysis, this study reveals five distinctive pathways of pilot sampling: *piloting for challenge*, *piloting for advancement*, *piloting for innovation*, *piloting for action*, and *piloting for regional generalization*. Each modality represents a specific purpose of experimental activities. A unique feature of this study is that the QCA results are substantiated with five illustrative case studies, thus synthesizing inductive and deductive approaches of reasoning. This study finds certain evidence for the positive sampling argument (Wang & Yang, 2021) in the healthcare arena, and suggests that the Chinese policymakers did consider a balance in terms of geographic and socioeconomic representativeness (He, 2022). More importantly, piloting serves as a versatile governance tool that can fulfill multiple functions in the context of complex reforms.

2 | POLICY EXPERIMENTATION AND PILOT SAMPLING

"*(P)olicy experimentation is different from controlled or lab-based experiments because often the political process of government involved in these experiments cannot be equated to the usual components of an experiment, such as having a hypothesis that can be tested through repeated trials, control groups, and randomization*" (Nair, 2020, p. 347). Political constraints, ethical considerations, or both often make it extremely hard to carry out randomized controlled trials in the world of public policy (Jowell, 2003; van der Heijden, 2014). Often used interchangeably with "policy experiment", policy pilot is not defined narrowly as a methodological approach based on a control-treatment group or randomization design––instead, it emphasizes the use of pilots as a tool for gaining relevant evidence and useful knowledge for policymaking in the face of vast uncertainties (Ko & Shin, 2017; Nair & Howlett, 2016).

Because randomization is rarely employed in reality, proper selection of pilots becomes even more important for the generation of robust and scalable policy lessons. Unfortunately, there is a paucity of knowledge about the strategic patterns of pilot sampling and their implications for experimental governance. A small number of studies have suggested that certain selection bias associated with pilots is a conducive factor for policy learning and scaling-up (Al-Ubaydli et al., 2019; DellaVigna & Kim, 2022; Gechter & Meager, 2021). Local authorities also have been found to compete for the pilot opportunity in many circumstances, because their participation is often seen as "doing a favor" for the national government (Ettelt et al., 2022; Zhu & Sun, 2009). In a major theoretical breakthrough, Ettelt et al. (2015) identified multiple purposes for piloting in the British welfare sectors, highlighting the value of piloting in promoting policy change and driving implementation.

China offers an excellent setting in which to unpack the black box of pilot sampling. Its long tradition of frequent policy experiments represents a hallmark of Chinese governance, contributing to problem-solving in a continent-size country with tremendous interregional disparities. In the Chinese language, "pilot" can refer to both an experimental project and an experimental site (Teets & Hasmath, 2020). In the well-known "experimentation under hierarchy" model proposed by Heilmann (2008), the central government defines the policy objectives, while local pilots are encouraged to search for appropriate policy instruments through widespread policy "tinkering" under an overarching national framework. Successful experiences gained from local pilots are subsequently synthesized to inform national-level policy formulation and some of them get scaled up to other regions. Even so, Heilmann (2008) did emphasize that local policy piloting in China should not be seen as "freewheeling trial and error or spontaneous policy diffusion"; instead, it is a "purposeful and coordinated activity geared to produce novel policy options that are injected into official policy making and then replicated on a larger scale". Ultimately, the central government is the body that determines the direction and process of experimentation, controls the experimental variables, and judges what constitutes the success of an experiment (Chan & Shi, 2022; Mei & Liu, 2014).

Notwithstanding its explanatory value, the model of "experimentation under hierarchy" overlooks the important issue of site selection. In a recent study, Zhu and Zhao (2021) suggest that when central policymakers do not have clear policy instruments at hand, they seem to pay greater attention to the *representativeness* of pilot sites, along geographic and socioeconomic lines. In the social welfare domain, especially healthcare, the Chinese government appears to be even more conscious of such a balance in order to generate solid experience that can inform policy synthesis at the national level (Barber et al., 2014; He, 2022; Zhu & Bai, 2020).

According to the standard protocol of a centrally steered experimental program, the State Council—or sometimes its ministries and commissions—kicks off the pilot selection process by

publishing general guidelines. Local governments then respond to the central guidelines with a local action plan, laying out the logistical and implementation details necessary for conducting the pilot (Wang & Yang, 2021). In many circumstances, pilot sites are selected directly by the central government on the basis of strategic considerations, but it should be noted that this process does not equate to a purely top-down assignment of mandatory work, because the central government must solicit local governments' willingness to participate (Barber et al., 2014; Chan & Shi, 2022; Zhou, 2013). This "collaborative work mode" makes sense because it is unwise to expect experimental success if the local authorities harbor strong reluctance. In other circumstances, the central government announces a call for applications, inviting interested local governments to bid for the pilot status. This typically happens when the experimental task is not too urgent and a fiscal package is attached. Indeed, in some policy domains, especially social welfare, local governments often lack the momentum to undertake pilots without additional fiscal support (Ettelt et al., 2022; Shi, 2012). Therefore, the central government may intentionally create a competition among local governments to bid for the pilot status and the associated fiscal package. Local governments that join such piloting through voluntary participation are typically recognized as having stronger motivation (Huang & Kim, 2020; Zhou, 2013). Even so, it must be underscored that in both mechanisms (assigned vs. voluntary), the ultimate decision of pilot sampling is made by the central government.

In a recent contribution, Wang and Yang (2021) analyzed a large number of policy pilots in China from the past 4 decades and revealed that more than 80% of them exhibited a clear mark of positive sample selection. In simple words, they found that the vast majority of pilots in China have been undertaken in wealthier regions, despite the professed goal of generating scalable solutions for the *entire* country. The authors explained this evident selection bias through misaligned incentives across China's political hierarchy, especially ministerial and local leaders' desire to gain a competitive edge in the "promotion tournament" through piloting. Yet, in another recent article that focused specifically on healthcare reforms, the central ministry was found to seek representativeness when selecting city-level pilots (He, 2022). In their study analyzing China's pilot program on long-term care insurance, Wang et al. (2022) noted that the central ministry selected those fiscally strong cities across the eastern, central, and western regions. In other words, the selection of pilots appeared to reflect both *representativeness* and *positive* sampling strategies. Overall, this inconsistency of results cited above suggests that the logic of pilot sampling is more complex than might be assumed.

3 | RESEARCH CONTEXT

Healthcare reforms are one of the most complex policy endeavors for governments. Numerous actors, institutions, informal rules, and deep material interests are intertwined and interact in nonlinear ways, making the outcome of any reform highly unpredictable (Husain, 2017). Therefore, an experimental approach is seen as a safe way to reform the health sector. The Chinese health policymakers were facing such a situation when the country's landmark national healthcare reform was launched in 2009. Three salient characteristics defined the context of that reform program. First, social health insurance was making an impressive expansion throughout the entire nation, providing the population with basic yet increasing financial protection against catastrophic illnesses. Even so, health expenditures were still escalating at a rather rapid rate, mainly driven by a plethora of misaligned incentives within the public hospital system. Expensive access to care remained a problem for many disadvantaged people (Barber et al., 2014).

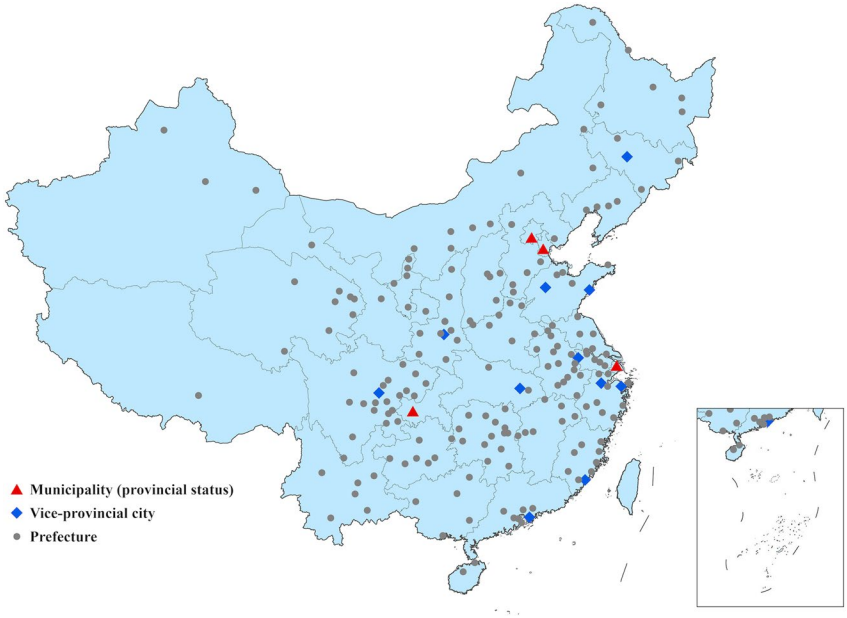

FIGURE 1 Geographic distribution of pilot cities. *Source*: drawn by authors.

Second, public hospitals, the chief provider of health services in China, were also the victims of the country's earlier, market-oriented health reforms. Fiscal subsidies to public hospitals drastically dwindled beginning in the 1980s, while the government instead allowed hospitals to earn revenues from patients in order to break even (Hsiao, 1995). Many perverse incentives were subsequently created by ill-designed policies that effectively transformed Chinese hospitals into profit-driven entities (Barber et al., 2014; He, 2011). Vast overprescribing of drugs and diagnostic tests was widespread across the country, creating an enormous waste of resources (Yip & Hsiao, 2008). Doctor-patient relationships deteriorated to an alarming level.

Third, health governance in China was very fragmented, as multiple ministries at the central level were involved in policymaking. As a result, central policymaking typically involved tedious attempts at coordination (Hsiao, 2007), and in the meantime, reform inertia was widely observed at the local level. For one thing, local governments expected clear reform guidelines from Beijing, which for its part barely reiterated broad policy objectives and offered no clear policy instruments. In addition, local governments were frequently found to take a passive stance toward reforms that did not carry a fiscal package (He, 2011). Public hospital reform was one such example.

Most of the consequences of the inappropriate incentives created in the past were reflected in public hospitals, creating a vicious deadlock. In consequence, public hospital reform was associated with extraordinary technical complexities. Previous local reforms had been mostly piecemeal in nature (Fu et al., 2017). Thus, by 2010, deep complexity and uncertainty prompted the central government to launch a large-scale nationwide experimental program that eventually listed 200 cities across the country in four phases. As is illustrated in Figure 1, the program covered virtually every provincial administrative unit of mainland China, and the pilot cities represented the various levels of administrative status. A wide range of experimental activities has been observed since then (Barber et al., 2014; He et al., 2022).

4 | RESEARCH DESIGN

4.1 | Methodology

Qualitative comparative analysis (QCA) is a genre of case-based and set-theoretic analysis developed by Charles Ragin (2009). Built on the logic of set theory and Boolean algebra, QCA can conduct systematic comparisons across cases and explore causal complexity by identifying multiple combinations of factors that are associated with the same outcome (Shephard et al., 2020). To investigate the selection of pilot sampling for the massive public hospital reform program in China, we assumed that a complex configurational causality was at play in the sense that the outcome of pilot selection was a function of the combined effects of a set of conditions from different dimensions. In other words, policymakers needed to consider various factors when selecting pilots (Wang & Yang, 2021; Zhou, 2013). Moreover, multiple sets of conditions forming various distinct types of cases could exist simultaneously to generate the same outcome from piloting. Thus, we considered QCA to be an appropriate analytical approach for this study.

When undertaking this research, two commonly used strategies—the crisp-set QCA (csQCA) and the fuzzy-set QCA (fsQCA)—were on our table. The former was adopted because dichotomization of conditions, as csQCA requires, is not merely a technique of simplification but well reflects real-world policy-making practices that often follow the binary logic of "do this or do that" (Blackman, 2013; Pattyn & Brans, 2015). Previous research has revealed that policymakers also expect information to be presented in a binary way that helps inform decision-making in complex systems (Soares et al., 2018). Moreover, fsQCA, despite its certain advantages, does not deliver the decisiveness of csQCA with binary conditions; the clarity of examining the combinations would become rather complicated too if ordinal or continuous values are used (Blackman, 2013).

A distinctive feature of our study was that each configuration generated through QCA was accompanied by a short, city-level case study in which the rationale of pilot selection and the uniqueness of such pilots were illustrated. Illustrative cases were selected from each configurational category based on the availability of secondary materials. This additional component strengthened our main results with necessary contextual richness.

4.2 | Case selection

The aim of this study was to explore the logic underlying the selection of pilot cities in China's public hospital reform program. As required by QCA, cases with a positive outcome (piloting) and those with a negative outcome (non-piloting) both needed to be included. Therefore, we intended to use all prefecture-level cities and municipalities in China as our sampling frame. The initial database included 315 cases, after removing cities from Xinjiang and Tibet due to low data availability. We also excluded 88 cities that had become pilots because they belonged to the 11 centrally-designated pilot provinces for the public hospital reform. The pilot provinces were selected in two rounds in 2015 and 2016 by the State Council, for the purpose of accelerating the reform, so those cases needed to be eliminated from the sample because their membership happened automatically and the presence of their outcomes could not be explained by any configuration of the conditions identified in this study. Last, we further excluded 18 cities due to a lack of data on health expenses per capita, which was a key explanatory condition. Ultimately, our sample consisted of 209 cases, including 103 pilots and 106 non-pilots. That sample size is commonly considered a large-*N* sample in QCA analyses (Greckhamer et al., 2013).

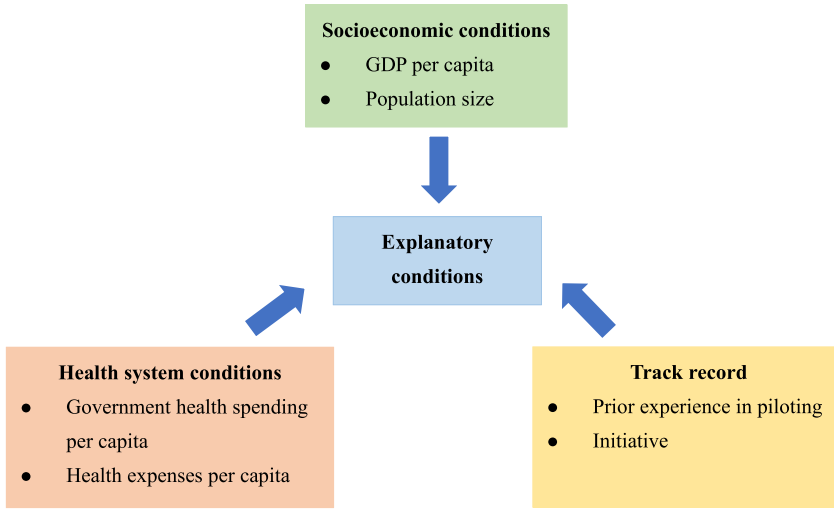

FIGURE 2 Explanatory conditions in qualitative comparative analysis.

4.3 | Outcome and explanatory conditions

The outcome of interest was enrollment in the pilot program. Synthesizing the received wisdom, we constructed three dimensions to capture the possible considerations in choosing a pilot sample: (1) socioeconomic conditions, (2) health system conditions, and (3) track record. For decisions about the number of explanatory conditions to be included, it is suggested as a good practice to keep a moderate number of conditions. The problem of limited diversity would occur if the number of theoretically possible configurations exceeds the number of cases, and too many conditions could also result in unnecessary complexity in interpreting the results and difficulty in generalizing patterns (Thomann & Maggetti, 2020). Because this study involved 209 cases, it allowed a maximum of seven explanatory conditions. After testing for various combinations of four to seven conditions, we eventually considered six conditions as most appropriate for our analysis. Each of the three dimensions included two conditions (see Figure 2).

The socioeconomic conditions dimension consisted of "GDP" (gross domestic product) per capita and "population size", with GDP per capita reflecting a city's economic strength that is positively associated with the distribution of health resources (Chen et al., 2021). The positive sampling argument by Wang and Yang (2021) was built primarily on economic capacity as represented by GDP per capita; we also posited that this factor was at play when central policymakers chose local pilots in public hospital reforms. Population size is traditionally measured by the number of permanent residents in a city. One important study noted that health policy experiments in China are more likely to happen in localities with larger populations (Wang & Yang, 2021), but other studies illustrated that pilots in health reforms tend to happen in localities with small to medium population sizes (Liu et al., 2002; Yang, 2022). The role of population size in pilot sampling thus warrants closer investigation. Taken together, we believed that these two conditions represented a city's basic economic and demographic conditions for undertaking pilots.

The health system conditions dimension was represented by government health spending per capita ("government spending"), and health expenses per capita ("health expenses"). The former was measured by the sum of total government spending on health divided by the number

of permanent residents in a city, and it enabled the comparison of a government's fiscal capacity and its willingness to spend on health across cases. Studies have found that cities with higher spending in the policy-specific categories are more likely to be selected as pilots because of their stronger fiscal capacity to make investments in piloting (Wang et al., 2022; Wang & Yang, 2021), and that governments with greater fiscal capacity show higher willingness to adopt innovative policies (Guo & Ba, 2020). If the positive sampling argument holds true, it is reasonable to speculate that policymakers are more likely to favor cities with stronger fiscal commitment in health when it comes to pilot selection. The condition of health expenses was used to gauge the financial burden of seeking care for local residents. Commonly referred to as "*kanbing gui*" (expensive access to care), this has been the thorniest health policy problem in China. Since the experimental program ultimately strived to seek effective solutions for this deep-seated problem, we posited that policymakers tend to choose cities where the financial burden to care is comparatively heavy, as the representativeness argument would suggest. Ideally, *kanbing gui* would be best measured by individuals' out-of-pocket spending, but unfortunately city-level data for this indicator was unavailable. We therefore adopted the percentage of health expenses in total living expenses of urban residents as a proxy, which can still reasonably reflect the general financial burden of seeking care in a city.

The track record dimension included two conditions: prior experience in piloting ("experience") and "initiative". Experience was defined as having a track record of undertaking national-level pilot in health sector reforms at least once prior to being selected into the current pilot program. If policymakers did conduct positive sampling, as we speculated, it was reasonable to assume that they were more likely to favor those with a track record, in order to secure experimental success. "Initiative" was defined as having made at least one locally initiated reform initiative in the health sector within the 5 years prior to a city's admission into the pilot. This condition was selected because local governments' willingness to participate in centrally steered reforms has been found to be crucial for ensuring desirable policy outcomes (He, 2022), and bottom-up initiatives made in recent years had demonstrated a city's momentum for undertaking health reforms. For our analysis, we determined that the local policy initiatives had to carry at least one of the following features: (1) triggering considerable change in the local health system; (2) producing substantial results; and (3) demonstrating originality and creativity.

4.4 | Data collection

All data were collected from secondary sources. Data on *GDP* per capita, *population size*, *government spending*, and *health expenses* were collected from official statistical yearbooks.[1] For pilot cities, these time-varying variables were lagged by 1 year to avoid simultaneity between the selection of pilots and changes in explanatory conditions. The data for time-varying variables for non-pilot cities were collected for the year 2015, because no more admission into the pilot programs was offered after that year.

Data on *experience* in piloting were obtained through web searches from Google, government websites, and the PKULaw policy document database. We first identified eligible pilot programs and the corresponding lists of pilot cities using search terms such as "national-level pilots," "pilot cities," and "health-sector reforms." The remaining cases were then searched on an individual basis. For pilot cases in this study, the experience was limited to previous pilots undertaken between 1989 and the year before the city's admission into the current pilot program. For non-pilot cases, the timespan was set between 1989 and 2015. The year 1989 was chosen as the starting point because it witnessed the start of major health sector reforms in China (WHO, 2015).

Data on *initiative* were extracted from media coverage. We systematically searched the Chinese National Knowledge Infrastructure database and newspaper database, particularly their collection of city-level official newspapers in which local policy innovations are often reported (Ma, 2017; Zhu & Zhang, 2019). Various search words, including "health sector reforms", "medical system reforms", "public hospitals", "institutional reforms", and "innovative practices in health services" were applied to locate relevant articles, and only those containing concrete information about reform initiatives were counted. For pilot cases, we searched for articles published within 5 years before the case was selected as a pilot. For non-pilot cases, we collected articles published between 2011 and 2015. Because the coding for *experience* and *initiative* entailed certain judgment calls on the part of the researchers, those terms were coded jointly by two coauthors to minimize bias. Data were first collected by one coauthor and then reviewed thoroughly by another. When disagreement emerged, three coauthors undertook further discussions and performed subsequent rounds of searches until a complete agreement was reached.

4.5 | Dichotomization of the data

The outcome condition and explanatory conditions were coded into quantitative membership scores to be analyzed by QCA. Because this study employed crisp-set QCA, each condition was coded to a binary membership score of either 1 (fully in the set) or 0 (fully out of the set). For the outcome of interest, pilot cities were assigned a score of 1, and 0 was assigned to non-pilot cities. For explanatory conditions, *GDP*, *government spending*, and *health expenses* were coded using the corresponding national-level data as thresholds. Cases with values greater than or equal to the thresholds were coded as 1, and otherwise were coded 0. Population size was coded according to the city classification standards released by the State Council in 2014. A score of 1 was assigned to cases with a population larger than or equal to 5 million, and 0 otherwise. For *experience* in piloting, cases took the value 1 if they had previously been selected as pilots at least once, and 0 otherwise. *Initiative* was coded using the same approach: Cases were given a membership score of 1 if reform initiatives were reported at least once, and 0 otherwise. Details of the operationalization process are reported in Appendix 2. Descriptive statistics for the explanatory conditions are presented in Table 1.

5 | ANALYSIS

The QCA analysis involved two key steps: a necessity analysis for identifying necessary conditions, followed by a sufficiency analysis for identifying sufficient configurations. All analyses were conducted using the R software 4.2.2 with packages QCA 3.18 (Duşa, 2019) and SetMethods 3.0 (Oana & Schneider, 2018). The necessity analysis determined whether any selected condition needed to be present for the outcome to occur (Mello, 2021). The degree of necessity was assessed by three measures: consistency, coverage, and relevance of necessity (RoN), with consistency referring to the proportion of cases displaying a given condition in the subset of cases presenting the outcome. If the consistency score reached 0.9, the condition could be deemed potentially necessary for the occurrence of the outcome (Oana et al., 2021). Such condition was then examined by its coverage and RoN. Coverage indicated the degree of empirical relevance of a condition that had been regarded as consistent and measured the proportion of cases exhibiting the outcome in the subset of cases with a given condition (Oana et al., 2021). The RoN evaluated

TABLE 1 Descriptive statistics of explanatory conditions.

	Pilots ($n = 103$)		Total ($n = 209$)	
Conditions	***n***	**%**	***n***	**%**
Socioeconomic conditions				
GDP	55	53	85	41
~GDP	48	47	124	59
POP	29	28	62	30
~POP	74	72	147	70
Health system conditions				
SPND	37	36	59	28
~SPND	66	64	150	72
EXPS	50	49	108	52
~EXPS	53	51	101	48
Track record				
EXP	63	61	94	45
~EXP	40	39	115	55
INIT	38	37	53	25
~INIT	65	63	156	75

Note: The tilde signs indicate absence of a condition.

Abbreviations: EXP, prior experience in piloting; EXPS, health expenses per capita; GDP, GDP per capita; INIT, initiative; POP, population size; SPND, government health spending per capita.

Source: Authors' data.

the trivialness of a condition, with lower values reflecting more trivialness and less relevance (Schneider & Wagemann, 2012). If both measures lay above 0.5, the potentially necessary condition could be regarded as relevant (Mello, 2021). As reported in Table 2, none of the conditions in our study showed consistency levels greater than or equal to 0.9. In other words, we did not identify any condition that cases absolutely had to fulfill in order to be selected as pilots. Apart from single necessary conditions, we further examined the SUIN conditions (combinations of conditions necessary for the outcome) and found no such combinations for this study. The results of the necessity analysis are visualized in Appendix 3.

Given that there were no necessary conditions in our study, we then performed a sufficiency analysis to identify sets of configurations that were always present when the desired outcome occurred. The results were presented by a "truth table", which transformed the membership scores into a list of all logically possible combinations with corresponding cases and associated outcomes (see Appendix 4 for details). The next step was to determine the sufficient combinations to be included in the subsequent analysis, using the selection criteria of consistency cutoff and frequency cutoff. This study applied a consistency threshold of 0.75 and a frequency threshold of 2, as has been recommended by scholarly conventions (Mello, 2021; Sager & Thomann, 2017). The analysis of "truth table" generated three types of solutions: complex, parsimonious, and intermediate solutions. We report the intermediate solution here as the main result.[2] The three solutions were originally produced via the Standard Analysis (SA). Yet, given the increasing popularity of the Enhanced Standard Analysis (ESA) that enabled the exclusion of untenable assumptions (Oana et al., 2021), we re-performed the analysis using ESA. The results of the intermediate solution remained identical.[3]

TABLE 2 Results of necessity analysis.

Conditions	Consistency	Coverage	RoN
GDP	0.53	0.65	0.81
~GDP	0.47	0.39	0.53
POP	0.28	0.47	0.82
~POP	0.72	0.50	0.46
SPND	0.36	0.63	0.87
~SPND	0.64	0.44	0.41
EXPS	0.49	0.46	0.64
~EXPS	0.51	0.52	0.69
EXP	0.61	0.67	0.79
~EXP	0.39	0.35	0.56
INIT	0.37	0.72	0.91
~INIT	0.63	0.42	0.37

Note: The tilde signs indicate absence of a condition.

Abbreviations: EXP, prior experience in piloting; EXPS, health expenses per capita; GDP, GDP per capita; INIT, initiative; POP, population size; RoN, relevance of necessity; SPND, government health spending per capita.

Source: Authors' data.

Table 3 reports the intermediate results and corresponding cases for each solution. Five pathways led to the outcome of being selected into the pilot program, and each pathway represented a unique combination of explanatory conditions. We characterize them as: piloting for challenge, piloting for advancement, piloting for innovation, piloting for action, and piloting for regional generalization. The core and peripheral conditions are distinguished in the table by the size of the icons. Core conditions were those present in both intermediate and parsimonious solutions, while peripheral conditions existed only in the intermediate ones. The core conditions were therefore more directly associated with the outcome than the peripheral ones were. All pathways showed a consistency score higher than 0.8, suggesting that each pathway was individually sufficient to generate the desired outcome (Greckhamer et al., 2018; Mello, 2021). The raw coverage represented the proportion of cases in the outcome set that could be explained by a certain pathway, whereas the unique coverage expressed the proportion of cases that could only be explained by a certain pathway.

In addition to examining individual solutions, our results also evaluated the reliability and validity of the five pathways, using two measures: solution consistency and solution coverage. Solution consistency measured the degree to which the complete set of solutions was consistent in producing the desired outcome. Generally, a minimum consistency value of 0.8 is recommended, regardless of the sample size (Greckhamer et al., 2018). As is reported in Table 3, the solution consistency of our result was 0.88, which is sufficient for a consistent subset relationship. Solution coverage assessed the empirical relevance of the overall result by measuring how well it represented the cases. The solution coverage of our main results was 0.54, suggesting that the five pathways jointly explained 54% of all the cases selected as pilots. There is no conventional threshold for solution coverage, as it is highly dependent on research design (Schneider & Wagemann, 2012). Typically, small-*N* QCA analyses are likely to attain a relatively high or nearly perfect solution coverage, whereas the values for large-*N* studies are usually lower (Greckhamer et al., 2013). Given our large sample size, a solution coverage of 54% is fairly high.

TABLE 3 Results of the intermediate solutions.

	S1	S2	S3	S4	S5
Socioeconomic conditions					
GDP per capita	●	●	●	⊗	⊗
Population size	⊗	⊗	•		⊗
Health system conditions					
Government spending	●	⊗		⊗	●
Health expenses	⊗			●	●
Track record					
Experience		●	●	●	●
Initiative			●	⊗	
Consistency	1	0.82	0.86	0.86	1
Raw coverage	0.10	0.17	0.12	0.12	0.04
Unique coverage	0.10	0.17	0.12	0.12	0.04
Solution consistency			0.88		
Solution coverage			0.54		
Typical cases	Xilingol, Wuhai, Xiamen, Yingtan, Dongying, Huizhou, Fangchenggang, Sanya, Panzhihua	Taiyuan, Hohhot, Baotou, Tongliao, Anshan, Panjin, Songyuan, Zhenjiang, Shaoxing, Wuhu, Xinyu, Jiaozuo, Zhuzhou, Zhuhai, Liuzhou, Haikou, Guiyang	Beijing, Tangshan, Shanghai, Hangzhou, Ningbo, Qingdao, Luoyang, Wuhan, Changsha, Shenzhen, Dongguan, Kunming	Handan, Yangquan, Tonghua, Qiqihaer, Shuangyashan, Pingxiang, Hebi, Puyang, Qujing, Dali, Baoji, Xining	Pu'er, Honghe, Jiuquan, Qingyang

Note: ● = core condition (present); ⊗ = core condition (absent); • = peripheral condition (present); ⊗ = peripheral condition (absent). No icon indicates that neither presence nor absence of the condition is important for the pathway.

Source: Authors' data.

We performed robustness checks to corroborate the main results, by using the methods developed by Oana et al. (2021). Omitted here due to page limit, the robust checks (reported in Appendix 9) confirmed the relative stability of the solutions.

6 | MULTIPLE PATHWAYS OF PILOT SAMPLING

6.1 | Solution #1: Piloting for challenge

The first configuration showed that high GDP per capita, high government health spending, low individual health expenses, and below-average population size jointly paved a city's way to being selected as a pilot, with the first three being core conditions. This configuration generally represented cases that had favorable economic conditions—which allowed the pilots to invest in reform—and small to medium population sizes with adequate health resources. The problem of *kanbing gui* was generally moderate. These features indicated that cases in this configuration had decent socioeconomic conditions for becoming a pilot, with necessary capacity also being present, thereby suggesting a greater chance of experimental success. By selecting such cities, policymakers aimed at encouraging new local initiatives to attack the health policy problems and to search for more scalable solutions. These pilots were expected to mobilize existing resources and unleash their potential to take on the reform's challenge and seek breakthroughs in public hospital reform. Therefore, we labeled them "piloting for challenge".

6.2 | Illustrative case #1: Xiamen

Located on China's southeast coast, Xiamen is the wealthiest city of Fujian Province but has a rather small population. Thanks to its enviable government coffer, government health spending per capita in Xiamen was almost twice the national average. Meanwhile, local residents' financial burden from seeking care was very low. These conditions put Xiamen in a favorable position to undertake public hospital reform. Prior to becoming a national pilot, Xiamen had already made significant progress in healthcare reforms. It was among the first few cities that built a universal health insurance system for both urban and rural residents. A pioneer electronic health record system was launched, covering 95% of health facilities in the city. Its high economic capacity enabled Xiamen to make substantive investments in public hospital reform, thereby allowing it to start the reform earlier than other pilots and to take greater actions. As committed by the Mayor, more funding would be allocated to public hospitals to support the pilot.[4] The rationale for the central government to select this city is vividly mirrored in the words of the Deputy Health Minister, who emphasized that Xiamen had gained rich experience in public hospital reforms and was capable of setting realistic policy goals aligned with the national blueprint. The Ministry of Health (MoH) vowed to continue to support Xiamen's pilot and promote scalable practices to other cities when appropriate.[5] He also underscored that Xiamen, as a pioneer in China's economic reforms, should be very innovation-minded and embrace great challenges in public hospital reform.[6] Another Deputy Minister also noted that while Xiamen had made satisfactory progress in the past few years, the central government expected the city to strive for even more challenging goals and produce more novel practices.[7]

6.3 | Solution #2: Piloting for advancement

The second configuration encompassed high GDP per capita, small to medium population size, low government health spending, and previous experience in piloting, all of which were core conditions. Similar to Solution #1, cases in this pathway embodied adequate economic capacity

and moderate population size, both pointing to an ideal basis for undertaking health reforms. Pilots in small regions tend to be considered less risky because an adverse impact would not be too catastrophic in the case of experimental failures (Jowell, 2003; Yang, 2022). Also important was the fact that these cities had previously conducted national-level pilots in health sector reforms, hence arguably leaving them with greater experience than their peers. Still, one notable drawback was their lack of government spending on health, which could lead to insufficient provision of resources in support of public hospital reform. Cities with these conditions were brought into the pilot program mainly for the purpose of fixing local problems while producing replicable lessons. Because policy pilots sometimes come with an injection of funds and other resources, the selection of pilots on this pathway could be seen as an opportunity for these cities to strengthen their capacity and seek solutions for local healthcare problems (Bailey et al., 2017; Ko & Shin, 2017). As a result, we named this configuration "piloting for advancement".

6.4 | Illustrative case #2: Zhuhai

Zhuhai is a medium-sized city in wealthy Guangdong Province. Its GDP per capita is far higher than the national average, but the city has the smallest population of the entire province. Despite its affluence, government health spending per capita in Zhuhai was actually lower than the national average. This city had previous experience being a pilot in healthcare reforms––in 1996, it was selected by the central government to participate in an experiment with fixed hospital payment mechanisms that was subsequently promoted in the province.[8] On the one hand, Zhuhai is a small city with economic strength, and was considered an ideal place for piloting. As the Deputy Director of the provincial health commission put it: "*Zhuhai has good experimental conditions (for the public hospital reform). It is not the wealthiest in the province (so there is a certain representativeness). There is no major social burden in the healthcare system, either (so it is less risky to undertake experiment).*" A senior official also noted that a moderate population size was conducive to pilot design and implementation, because "a small boat turns around more easily."[9] On the other hand, a lack of government spending on health had resulted in insufficient funding to support public hospitals, most of which relied on overprescribing to break even. It was estimated that government funding alone could only maintain 15 days of operation for big local public hospitals in this city.[10] Following its admission into the national pilot, Zhuhai focused its reform on two critical fronts: normalizing drug pricing and adjusting the hospital financing system, both of which were not only acute local problems but were of nationwide importance.[11] Thus, Zhuhai's selection into the pilot provided a good opportunity for the city to seek solutions to local health problems and make advancement, as the Deputy Health Minister noted.[12]

6.5 | Solution #3: Piloting for innovation

The third pathway combined high GDP per capita with a large population size, experience in piloting, and a strong initiative. Large population size was a peripheral condition, whereas the rest constituted core conditions. A dominant majority of cities in this solution were either the wealthiest ones in China (e.g., Beijing, Shanghai, Shenzhen) or provincial capitals (e.g., Hangzhou, Wuhan, Changsha), which shared the commonalities of a sizable population and economic strength. Although such cities generally face greater difficulties in engineering large-scale health system reforms, due to the high stakes and complexities, they have displayed strong initiatives

for generating novel practices in previous years, accompanied by experience in undertaking national-level pilots. Those attributes enabled them to overcome the barriers encountered during the reform and to promote innovative local practices that could inform national policymaking. The government's main purpose in selecting this group of pilots was to create an opportunity to initiate local change as the first step toward wider scaling-up (Ettelt et al., 2015; Wang et al., 2022). Consequently, this pathway was characterized as "piloting for innovation".

6.6 | Illustrative case #3: Shanghai

As one of the most prosperous cities in China, Shanghai is characterized by a large population and a well-developed health system. It has a rich record of undertaking national health reform pilots, also having made several local initiatives prior to the national public hospital reform. A significant example of those initiatives was the corporatization of public hospitals in 2005, which was lauded as a successful innovation by the central government. Representing a breakthrough in hospital governance reform, Shanghai's "Shen Kang model" inspired many local governments in restructuring their hospital systems. That combination of characteristics made Shanghai an ideal candidate for the new pilot. A senior official of the MoH stressed that strong local initiative was a critical consideration for national pilots, and cities with prior voluntary initiatives were more likely to become national pilots because of their favorable conditions for reform.[13] The Deputy Health Minister also indicated that the decision to designate Shanghai as a national pilot was made in recognition of its remarkable achievements in this regard.[14] As underlined by another ministerial official, the central government expected Shanghai to provide additional novel practices for national scaling-up through this pilot.[15]

6.7 | Solution #4: Piloting for action

In contrast to Solutions #1 through 3, with their high GDP per capita, the fourth pathway comprised below-average GDP per capita, low government spending, high health expenses, previous experience in piloting, and a low level of initiative. All conditions except low government health spending were core elements. These conditions suggest that cities on this pathway suffered from salient problems in their health systems but lacked strong momentum for reform. On the one hand, this group of cities were less economically developed, with low government health spending and a comparatively high financial burden of healthcare on local residents. Furthermore, they seemed not to have demonstrated strong motivation to embark on healthcare reforms in the past. We propose that policymakers selected these cases for pilots primarily with the purpose of imposing a certain pressure to stimulate local change. In other words, piloting could be understood as a tool for central policymakers to "steer at a distance" and to construct a mandate for localities to tackle thorny problems (Ettelt et al., 2022). Thus, we labeled this pathway "piloting for action".

6.8 | Illustrative case #4: Baoji

Located in the northwestern hinterland, Baoji is a medium-sized city in Shaanxi Province. Although its general level of economic status is in the province's middle stratum, Baoji's GDP per

capita still falls far below the national average. Government spending on health was relatively low, while people's financial burden for seeking care was quite high. The city had been selected by the central government in 2007 as a pilot for the Urban Residents Basic Health Insurance Scheme. Its limited economic capacity greatly hindered the local government from adequate investment in healthcare reforms. Since 2003, government funding for public hospitals had drastically dwindled. A budget of 50 million was committed in 2010 to support the reform, but that was eventually reduced to 20 million.[16] As the Deputy Mayor admitted, cities in the western regions were faced with grave difficulties in funding such reforms, and financial assistance from the central government was desperately needed.[17] These obstacles led to a lack of momentum for undertaking the new pilot, and in fact, Baoji did not apply for pilot status––its application was actually mandated by the provincial government.[18] However, its inclusion into the national pilot program generated necessary pressure for the local government, which subsequently worked out a pragmatic pilot plan well suited to local needs. As the Vice Governor reiterated, all government departments of Baoji should be aware of the complexities associated with the public hospital reform and should take concrete actions in the face of such difficulties.[19]

6.9 | Solution #5: Piloting for regional generalization

In the fifth pathway, low GDP per capita, small to medium population size, high government health spending, high individual health expenses, and previous experience in piloting combined to qualify cities for inclusion in the national pilot program. All but small to medium population size were core conditions. This pathway covered the lowest number of cases, and they included four small to medium-sized cities in Yunnan and Gansu, two of the most underdeveloped provinces in China. As with Solution #4, this pathway exhibited a weak economic capacity and a severe problem of medical impoverishment. However, it was also marked by relatively sufficient government spending on healthcare (vis-à-vis other poor peers), in addition to having prior experience in piloting––both of which represented useful advantages in undertaking public hospital reforms. We therefore suggest that cases of this route resembled those that had stronger track records than other comparable cities in the underdeveloped western regions. These four cities were selected with the hope of testing the feasibility of national policies in less-developed contexts and of producing useful solutions that would be replicable in other underdeveloped regions. Thus, we characterized this solution as “piloting for regional generalization”.

6.10 | Illustrative case #5: Qingyang

Located in the northwestern Gansu Province, Qingyang lags far behind the rest of the country in economic development. Local residents' financial burden in seeking care was twice the national average. Notably, however, the per capita government spending on health in Qingyang was slightly higher than the national average. Qingyang also had previous experience as a health reform pilot, as it was one of the national pilots for the first phase of China's national essential medicines reform in 2010. Overall, this city had relatively satisfactory conditions for serving as a pilot, vis-à-vis its peers in underdeveloped western China. In 2011, Qingyang was first selected as a provincial pilot for public hospital reform. A series of efforts were made after that, such as increasing government funding and the level of insurance benefits, strengthening the capacity of primary care, and so on. This city was explicitly given the mission of producing useful practices

that could be scaled up in Gansu Province.[20] Its performance was recognized by the provincial government as "paving the way for provincial reform". The provincial government expected Qingyang to generate additional scalable practices in the future. [21]

7 | DISCUSSION AND CONCLUSIONS

Experiments in the policy world hardly follow the rules for laboratory settings (Ko & Shin, 2017; van der Heijden, 2014). Governments do enjoy a great deal of discretion in setting the terms and conditions of piloting design (Jowell, 2003; Nair, 2020). The selection of pilot sites is arguably one the most fundamental decisions to be made by policymakers, but unfortunately, scholars know little about the logic behind such decisions. Two competing theoretical explanations prevail. The representativeness argument suggests that policymaker *should* and *do* pay strategic attention to the socioeconomic, fiscal, and geographic representativeness of pilot sites in order to maximize the nationwide usefulness of experimental lessons (He, 2022; Zhu & Zhao, 2021). The positive sampling argument, however, contends that policymakers, driven by myriads of motivations, intentionally select localities with favorable conditions to undertake pilots (Wang & Yang, 2021).

The empirical evidence gleaned thus far appears contradictory. Our study attempted to resolve this paradox with the case of China's experimental program in public hospital reforms. We adopted a novel research design by combining the QCA method with illustrative cases to elucidate the logic of selecting pilots. In particular, QCA enabled us to unpack the rich causal dynamics behind such complex policy decisions in a large governance system. Three key insights arose.

First, pilot sampling is by no means a linear decision––multiple logics are at play. We found moderate evidence to support the positive sampling argument, because just 53% (55 out of 103) pilot cities in our study enjoyed above-average economic status as represented by their GDP per capita. The rationale for positive sampling is not difficult to understand due to the nature of the pilots: many policy reforms in the welfare sectors require fiscal commitment. High government spending on health also constituted a key condition in two out of the five solutions we identified. In the meantime, however, our results also found considerable support for the representativeness argument. Some pilot cities were neither economically better-off ones nor major cities (especially in Solutions #4 and #5), but they confronted grave local healthcare problems, especially cost escalation and the resultant expensive access to care, thus mirroring the perverse incentives deeply embedded in their respective public hospital systems. The addition of these cities into the program appears to have been driven by a keen desire to solve local problems. Policymakers also demonstrated a high expectation for these cities to produce regionally generalizable experiences through serving as pilots, as was most evident in Solution #5.

Second, we actually found another peculiar form of positive sampling in the sense that a vast majority of the pilot cities had prior experience in health sector reforms. In fact, the condition of *experience* represented a core condition in all but one solution. It is easy to understand that localities with track records in domain-specific reforms would generally be expected to offer a brighter prospect as pilots. It would be even better if the current piloting were built on their existing reform initiatives, so that policy continuity could strengthen the lessons learned. In the meantime, though, we contend that this significant presence of experience in pilot sampling may reflect a certain form of cherry-picking behavior that is not entirely conducive to optimal experimental outcomes. Operating in complex sociopolitical environment, policymakers are driven by career incentives when designing policy experiments. It is unsurprizing that designating pilots

in regions with rich reform experience can help increase the chances of experimental success, but such a sampling bias might also lead to disproportionate underrepresentation of "laggards" where entrepreneurial reforms are probably needed more.

Last, yet also importantly, this study echoes other recent ones (Ettelt et al., 2015; He, 2022; He et al., 2022; Nair & Howlett, 2016) that found the use of pilots to be a versatile governance tool capable of serving multiple purposes. The five distinctive pathways elucidated through our study manifested the rich dynamics of policy experimentation. The possible considerations of policymakers in the sampling stage in fact correspond to specific experimental purposes, such as challenge, innovation, generalization, and so forth, as we have labeled them. Whereas some pilots are given explicit expectations for trialling nationally useful innovations, others are intended primarily to solve local problems and hold only modest ambitions for generalization. We also noted that piloting can be used as a tool to stimulate local efforts, because the status of being a pilot generates necessary expectations and pressures on the implementers. All these purposes are not mutually exclusive but in fact vividly mirror the rich governance dynamics faced by policymakers in large countries.

This study is certainly not with limitations. Importantly, a stronger research design to probe the ultimate purposes of pilot selection would inevitably require a qualitative study supported by in-depth elite interviews with senior policymakers. The condition of doing so was however unavailable to us when this research was being undertaken. In our future research, we seek to gain deeper understanding on the multifaceted logic underlying pilot selection through greater use of in-depth qualitative methods.

ACKNOWLEDGMENTS

This study benefits from the authors' multiple discussions with the following persons: Chunxiao Wang, Jiwei Qian, Yunpeng Song, Zhuang Cao, Hongqiao Fu, Liqiang Zhang, and Lele Li. This study was funded by the General Research Fund (GRF) of the Research Grants Council (Ref. 18605720), Hong Kong SAR.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data used in the comparative qualitative analysis (QCA) is available upon request. The authors are pleased to share the dataset in the form of online supplementary file should this paper get accepted for publication.

FILE DESCRIPTION

This file contains full dataset used for QCA analysis, including case names, outcome of interest, and calibrated data of explanatory conditions. The 209 cases include 103 pilots and 106 non-pilots. The explanatory conditions consist of GDP per capita, population size, government health spending per capita, health expenses per capita, prior experience in piloting, and initiative.

ORCID


Alex Jingwei He https://orcid.org/0000-0001-9024-4831

ENDNOTES

[1] The descriptive statistics of time-varying raw data are reported in Appendix 1.

[2] The intermediate solution is generated by including logical remainders consistent with substantive knowledge into the minimization process. This is in contrast with the complex solution that only accepts empirical observations and the parsimonious solution that considers all logical remainders. This study yields identical results for the complex and intermediate solutions because existing theoretical or substantive knowledge cannot lead us to make directional expectations about the relationship between our conditions and the outcome. As a rule of thumb, the intermediate solution is reported here. Results of the other two solutions are included in Appendices 5–8.

[3] The ESA is advantageous in identifying logical remainder rows that lead to untenable assumptions and ruling them out from the logical minimization process. The untenable assumptions are classified into three types: (1) incoherent counterfactuals contradicting the statement of necessity; (2) incoherent counterfactuals contradicting assumptions; and (3) implausible counterfactuals contradicting common sense (Oana et al., 2021). Given that no necessary conditions were present and any combinations of the selected conditions were logically plausible, we only excluded logical remainder rows that contradicted simplifying assumptions. Four such rows (26, 28, 58, and 59) were identified using the methods developed by Oana et al. (2021), and were excluded from the minimization of the outcome "enrollment in the pilot program". The enhanced solution formulas were then produced based on the updated truth table.

[4] Liaowang Weekly, "*The performance of three public hospital reform pilots,*" March 15, 2010 (in Chinese), retrieved from https://news.sina.com.cn/c/sd/2010-03-15/112819867010.shtml.

[5] Ministry of Health, "*Deputy Ministry of Health visits Xiamen to oversee public hospital reform pilot,*" April 22, 2011 (in Chinese), retrieved from http://www.gov.cn/gzdt/2011-04/22/content_1850415.htm.

[6] Ibid.

[7] China National Radio, "*Ministry of Health delegation visits Xiamen to oversee the public hospital reform pilot,*" August 23, 2011 (in Chinese), retrieved from https://news.ifeng.com/c/7faDT3AkeIW.

[8] Zhuhai Daily, "*Interview with Zhang Xiaotian: the pathfinder of the health insurance system of Zhuhai,*" September 10, 2020 (in Chinese), retrieved from http://zhuhaidaily.hizh.cn/html/2020-09/10/content_1212_3267631.

[9] 21st Century Business Herald, "*The triangle model of health reform in Zhuhai,*" April 10, 2009 (in Chinese).

[10] Xin Shiji (New Age), "*The fall of public hospitals in Zhuhai,*" January 3, 2021 (in Chinese), retrieved from https://magazine.caixin.com/2011-01-01/100213253.html.

[11] Zhuhai Daily, "*The public hospital reform of Zhuhai to be launched on March 29,*" March 16, 2015 (in Chinese), retrieved from http://med.china.com.cn/content/pid/15656/tid/3.

[12] Zhuhai Health Commission, "*National health official approves health reform progress of Zhuhai,*" June 12, 2016 (in Chinese), retrieved from http://wsjkj.zhuhai.gov.cn/zwgk/gzdt/content/post_2059120.html.

[13] 21st Century Business Herald, "*Prior experience will be taken into account when selecting health reform pilots,*" March 25, 2008 (in Chinese), retrieved from http://news.sina.com.cn/c/2008-03-25/011315215034.shtml.

[14] Ministry of Health, "*Ma Xiaowei visits Shanghai to oversee the public hospital reform pilot,*" March 5, 2010 (in Chinese), retrieved from http://www.gov.cn/gzdt/2010-03/05/content_1548735.htm.

[15] Ministry of Health, "*Shanghai held promotion conference for deepening the public hospital reform,*" December 28, 2012 (in Chinese), retrieved from http://www.nhc.gov.cn/tigs/s3582/201212/a4c3de4c268242d7b9f15e74c8e4e814.shtml.

[16] Time Weekly, "*An investigation of the public hospital reform in Baoji,*" July 29, 2010 (in Chinese), retrieved from https://www.time-weekly.com/post/9281.

[17] See Endnote #5.

[18] See Endnote #16.

[19] Baoji Daily, "*Promote public hospital reform to benefit residents,*" December 23, 2009 (in Chinese).

[20] Longdong Daily, "*Officials of Gansu Provincial Health Commission visit Qingyang to investigate public hospital reform,*" March 21, 2010 (in Chinese).

[21] Gansu Provincial Hospital of Traditional Chinese Medicine, "*Deputy Director of Gansu Provincial Health Commission visits Qingyang and Pingliang to oversee public hospital reform,*" May 7, 2012 (in Chinese), retrieved from https://www.gszyy.com/Item/6077.aspx; Gansu Provincial Hospital of Traditional Chinese Medicine, "*Duan Wei investigates health and family planning work of Qingyang,*" June 19, 2015 (in Chinese), retrieved from https://www.gszyy.com/Item/12010.aspx.

REFERENCES

Al-Ubaydli, O., Lee, M. S., List, J. A., Mackevicius, C., & Suskind, D. (2019). How can experiments play a greater role in public policy? Twelve proposals from an economic model of scaling. *Behavioural Public Policy*, *5*(1), 2–49. https://doi.org/10.1017/bpp.2020.17

Ansell, C. K., & Bartenberger, M. (2016). Varieties of experimentalism. *Ecological Economics*, *130*, 64–73. https://doi.org/10.1016/j.ecolecon.2016.05.016

Bailey, S., Checkland, K., Hodgson, D., McBride, A., Elvey, R., Parkin, S., & Pierides, D. (2017). The policy work of piloting: Mobilising and managing conflict and ambiguity in the English NHS. *Social Science and Medicine*, *179*, 210–217. https://doi.org/10.1016/j.socscimed.2017.02.002

Barber, S. L., Borowitz, M., Bekedam, H., & Ma, J. (2014). The hospital of the future in China: China's reform of public hospitals and trends from industrialized countries. *Health Policy and Planning*, *29*(3), 367–378. https://doi.org/10.1093/heapol/czt023

Blackman, T. (2013). Rethinking policy-related research: Charting a path using qualitative comparative analysis and complexity theory. *Contemporary Social Science*, *8*(3), 333–345. https://doi.org/10.1080/21582041.2012.751500

Chan, W. K., & Shi, S. J. (2022). "Central coordination, regional competition, and local protectionism" social decentralization in China's long-term care reform. *Social Policy and Administration*, *56*(6), 956–969. https://doi.org/10.1111/spol.12854

Chen, J., Lin, Z., Li, L. A., Li, J., Wang, Y., Pan, Y., Xiao, L., Xu, C., Zeng, X., & Xie, X. (2021). Ten years of China's new healthcare reform: A longitudinal study on changes in health resources. *BMC Public Health*, *21*(1), 1–13. https://doi.org/10.1186/s12889-021-12248-9

DellaVigna, S., & Kim, W. (2022). *Policy diffusion and polarization across US states*. (NBER Working Paper No. 30142). National Bureau of Economic Research.

Duşa, A. (2019). *QCA with R: A comprehensive resource*. Springer.

Ettelt, S., Mays, N., & Allen, P. (2015). The multiple purposes of policy piloting and their consequences: Three examples from national health and social care policy in England. *Journal of Social Policy*, *44*(2), 319–337. https://doi.org/10.1017/s0047279414000865

Ettelt, S., Williams, L., & Mays, N. (2022). National policy piloting as steering at a distance: The perspective of local implementers. *Governance*, *35*(2), 385–401. https://doi.org/10.1111/gove.12589

Fu, H., Li, L., Li, M., Yang, C., & Hsiao, W. (2017). An evaluation of systemic reforms of public hospitals: The Sanming model in China. *Health Policy and Planning*, *32*(8), 1135–1145. https://doi.org/10.1093/heapol/czx058

Gechter, M., & Meager, R. (2021). *Combining experimental and observational studies in meta-analysis: A debiasing approach*. Pennsylvania State University. retrieved from https://www.personal.psu.edu/mdg5396/MGRM_Combining_Experimental_and_Observational_Studies.pdf

Greckhamer, T., Furnari, S., Fiss, P. C., & Aguilera, R. V. (2018). Studying configurations with qualitative comparative analysis: Best practices in strategy and organization research. *Strategic Organization*, *16*(4), 482–495. https://doi.org/10.1177/1476127018786487

Greckhamer, T., Misangyi, V. F., & Fiss, P. (2013). The two QCAs: From a small-N to a large-N set-theoretic approach. In P. C. Fiss, B. Cambré, & A. Marx (Eds.), *Configurational theory and methods in organizational research* (pp. 49–75). Emerald.

Guo, L., & Ba, Y. (2020). Adopt or not and innovation variation: A dynamic comparison study of policy innovation and diffusion mechanisms. *Journal of Comparative Policy Analysis*, *22*(4), 298–319. https://doi.org/10.1080/13876988.2019.1603603

He, J. A. (2011). China's ongoing public hospital reform: Initiatives, constraints and prospect. *Journal of Asian Public Policy*, *4*(3), 342–349. https://doi.org/10.1080/17516234.2011.630228

He, J. A. (2022). Scaling-up through piloting: Dual-track provider payment reforms in China's health system. *Health Policy and Planning*, *38*(2), 218–227. https://doi.org/10.1093/heapol/czac080

He, J. A., Fan, Y., & Su, R. (2022). Seeking policy solutions in a complex system: Experimentalist governance in China's healthcare reform. *Policy Sciences*, *55*(4), 755–776. https://doi.org/10.1007/s11077-022-09482-2

Heilmann, S. (2008). From local experiments to national policy: The origins of China's distinctive policy process. *The China Journal*, *59*, 1–30. https://doi.org/10.1086/tcj.59.20066378

Hsiao, W. C. (1995). The Chinese health care system: Lessons for other nations. *Social Science and Medicine*, *41*(8), 1047–1055. https://doi.org/10.1016/0277-9536(94)00421-o

Hsiao, W. C. (2007). The political economy of Chinese health reform. *Health Economics, Policy and Law*, *2*(3), 241–249. https://doi.org/10.1017/s1744133107004197

Huang, X., & Kim, S. E. (2020). When top-down meets bottom-up: Local adoption of social policy reform in China. *Governance*, *33*(2), 343–364. https://doi.org/10.1111/gove.12433

Husain, L. (2017). Policy experimentation and innovation as a response to complexity in China's management of health reforms. *Globalization and Health*, *13*(1), 1–13. https://doi.org/10.1186/s12992-017-0277-x

Jowell, R. (2003). *Trying it out: The role of 'pilots' in policy-making*. UK Cabinet Office. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/498256/Trying_it_out_the_role_of_pilots_in_policy.pdf

Ko, K., & Shin, K. (2017). How Asian countries understand policy experiment as policy pilots? *Asian Journal of Political Science*, *25*(3), 253–265. https://doi.org/10.1080/02185377.2017.1360784

Li, L., & Fu, H. (2017). China's health care system reform: Progress and prospects. *The International Journal of Health Planning and Management*, *32*(3), 240–253. https://doi.org/10.1002/hpm.2424

Liu, G. G., Zhao, Z., Cai, R., Yamada, T., & Yamada, T. (2002). Equity in health care access to: Assessing the urban health insurance reform in China. *Social Science and Medicine*, *55*(10), 1779–1794. https://doi.org/10.1016/s0277-9536(01)00306-9

Ma, L. (2017). Site visits, policy learning, and the diffusion of policy innovation: Evidence from public bicycle programs in China. *Journal of Chinese Political Science*, *22*(4), 581–599. https://doi.org/10.1007/s11366-017-9498-3

Mei, C., & Liu, Z. (2014). Experiment-based policy making or conscious policy design? The case of urban housing reform in China. *Policy Sciences*, *47*(3), 321–337. https://doi.org/10.1007/s11077-013-9185-y

Mello, P. A. (2021). *Qualitative comparative analysis: An introduction to research design and application*. Georgetown University Press.

Nair, S. (2020). Designing policy pilots under climate uncertainty: A conceptual framework for comparative analysis. *Journal of Comparative Policy Analysis*, *22*(4), 344–359. https://doi.org/10.1080/13876988.2019.1695973

Nair, S., & Howlett, M. (2016). Meaning and power in the design and development of policy experiments. *Futures*, *76*, 67–74. https://doi.org/10.1016/j.futures.2015.02.008

Oana, I. E., & Schneider, C. Q. (2018). SetMethods: An add-on R package for advanced QCA. *The R Journal*, *10*(1), 507–533. https://doi.org/10.32614/rj-2018-031

Oana, I. E., Schneider, C. Q., & Thomann, E. (2021). *Qualitative comparative analysis using R: A beginner's guide*. Cambridge University Press.

Pattyn, V., & Brans, M. (2015). Organizational analytical capacity: Policy evaluation in Belgium. *Policy and Society*, *34*(3–4), 183–196. https://doi.org/10.1016/j.polsoc.2015.09.009

Ragin, C. C. (2009). *Redesigning social inquiry: Fuzzy sets and beyond*. University of Chicago Press.

Sabel, C. F., & Zeitlin, J. (2010). *Experimentalist governance in the European Union: Towards a new architecture*. Oxford University Press.

Sager, F., & Thomann, E. (2017). Multiple streams in member state implementation: Politics, problem construction and policy paths in Swiss asylum policy. *Journal of Public Policy*, *37*(3), 287–314. https://doi.org/10.1017/s0143814x1600009x

Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. Cambridge University Press.

Shephard, D. D., Ellersiek, A., Meuer, J., Rupietta, C., Mayne, R., & Cairney, P. (2020). Kingdon's multiple streams approach in new political contexts: Consolidation, configuration, and new findings. *Governance*, *34*(2), 523–543. https://doi.org/10.1111/gove.12521

Shi, S. J. (2012). Social policy learning and diffusion in China: The rise of welfare regions? *Policy and Politics*, *40*(3), 367–385. https://doi.org/10.1332/147084411x581899

Soares, M. B., Alexander, M., & Dessai, S. (2018). Sectoral use of climate information in Europe: A synoptic overview. *Climate Services*, *9*, 5–20. https://doi.org/10.1016/j.cliser.2017.06.001

Teets, J. C., & Hasmath, R. (2020). The evolution of policy experimentation in China. *Journal of Asian Public Policy*, *13*(1), 49–59. https://doi.org/10.1080/17516234.2020.1711491

Thomann, E., & Maggetti, M. (2020). Designing research with qualitative comparative analysis (QCA): Approaches, challenges, and tools. *Sociological Methods and Research*, *49*(2), 356–386. https://doi.org/10.1177/0049124117729700

van der Heijden, J. (2014). Experimentation in policy design: Insights from the building sector. *Policy Sciences*, *47*(3), 249–266. https://doi.org/10.1007/s11077-013-9184-z

Wang, S., & Yang, D. Y. (2021). *Policy experimentation in China: The political economy of policy learning*. (NBER Working Paper No. 29402). National Bureau of Economic Research. Retrieved from: https://www.nber.org/papers/w29402

Wang, Y., Ma, L., & Christensen, T. (2022). The hybridization of policy learning in multilevel regimes: A case study of the long-term care insurance in China. *Journal of Asian Public Policy*, 1–25. https://doi.org/10.1080/17516234.2022.2103392

World Health Organization. (2015). People's Republic of China health system review. Retrieved from: https://apps.who.int/iris/handle/10665/208229

Yang, Y. (2022). The fable of policy entrepreneurship? Understanding policy change as an ontological problem with critical realism and institutional theory. *Policy Sciences*, *55*(3), 573–591. https://doi.org/10.1007/s11077-022-09463-5

Yip, W., & Hsiao, W. C. (2008). The Chinese health system at a crossroads. *Health Affairs*, *27*(2), 460–468. https://doi.org/10.1377/hlthaff.27.2.460

Zhou, W. (2013). *Study on China's policy piloting*. Tianjin People's Press. [in Chinese].

Zhu, X., & Bai, G. (2020). Policy synthesis through regional experimentations: Comparative study of the new cooperative medical scheme in three Chinese provinces. *Journal of Comparative Policy Analysis*, *22*(4), 320–343. https://doi.org/10.1080/13876988.2020.1700664

Zhu, X., & Sun, B. (2009). Tianjin Binhai new area: A case study of multi-level streams model of Chinese decision-making. *Journal of Chinese Political Science*, *14*(2), 191–211. https://doi.org/10.1007/s11366-009-9048-8

Zhu, X., & Zhang, Y. (2019). Diffusion of marketization innovation with administrative centralization in a multilevel system: Evidence from China. *Journal of Public Administration Research and Theory*, *29*(1), 133–150. https://doi.org/10.1093/jopart/muy034

Zhu, X., & Zhao, H. (2021). Experimentalist governance with interactive central-local relations: Making new pension policies in China. *Policy Studies Journal*, *49*(1), 13–36. https://doi.org/10.1111/psj.12254

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: He, A. J., Fan, Y., & Su, R. (2023). Unpack the black box of pilot sampling in policy experimentation: A qualitative comparative analysis of China's public hospital reform. *Governance*, 1–22. https://doi.org/10.1111/gove.12804