

The Impact of a Pay-for-Performance Scheme on Prescription Quality in Rural China

Impact Evaluation

Xiaojie Sun

Xiaoyun Liu

Qiang Sun

Winnie Yip

Adam Wagstaff

Qingyue Meng

The World Bank
Development Research Group
Human Development and Public Services Team
May 2014



Abstract

In China, health care providers have traditionally been paid fee-for-service and overprescribing and high out-of-pocket spending are common. In this study, township health centers in two counties were assigned almost randomly to two groups: in one, fee-for-service was replaced by a global capitated budget; in the other, by a mix of global capitated budget and pay-for-performance. Performance captured inter alia “irrational” drug prescribing; 20 percent of the global capitated budget was withheld each quarter, points were deducted for failure to meet targets, and some of the withheld budget was returned in line with the points deducted. Outcomes included appropriate prescribing and prescription cost,

data on which were obtained by digitizing prescriptions from a month just before the reform and from the same month a year later. Impacts were assessed via multivariate differences-in-differences with township health center fixed effects. To reduce bias from non-randomness in assignment, the sample was trimmed by coarsened exact matching. Pay-for-performance reduced inappropriate prescribing significantly and substantially in the county where the initial level was above the penalty threshold, but end-line rates were still appreciable; no effects were seen in the county where initial levels were around or below the threshold, or on out-of-pocket spending in either county.

This paper is a product of the Human Development and Public Services Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The authors may be contacted at awagstaff@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

The Impact of a Pay-for-Performance Scheme on Prescription Quality in Rural China: An Impact Evaluation

Xiaojie Sun^a, Xiaoyun Liu^b, Qiang Sun^a,
Winnie Yip^c, Adam Wagstaff^d, and Qingyue Meng^b

^a Center for Health Management and Policy (Key Lab of Health Economics and Policy, National Health and Planning Commission), Shandong University, Jinan, China

^b China Center for Health Development Studies, Peking University, Beijing, China

^c Blavatnik School of Government, University of Oxford, UK

^d Development Research Group, The World Bank, Washington DC, USA

JEL code: I1

Keywords: China; Pay-for-performance; Drug prescribing

Acknowledgements

The research reported in the paper was supported by a grant from the European Union's FP7 program to the "Health Equity and Financial Protection in Asia" (HEFPA) project (HEALTH-F2-2009-223166), and by the World Bank's Strategic Impact Evaluation Fund (SIEF) (TF094336). We are grateful to staff working in Junan Health Bureau and Liangshan Health Bureau of Shandong Province for support and technical advice on the intervention design and implementation; to Dr. Shasha Yuan, Dr. Zhiyuan Hou and other participants from Shandong University and Peking University for their help in designing and fielding the survey and in organizing and analyzing the data; to local participants in two project counties for their efforts in implementing and evaluating this intervention; and to participants at HEFPA workshops for comments on our work. The findings, interpretations and conclusions expressed in this paper are entirely those of the authors, and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments of the countries they represent, including that of China.

Introduction

Recent years have seen a growth of momentum toward universal health coverage (UHC) – the idea that everyone should be able to get necessary care without causing themselves and their family undue financial hardship (World Health Organization 2010; Savedoff et al. 2012). Much of the focus of policy makers has been on extending insurance coverage, by bringing in hitherto uncovered groups into existing insurance or pre-payment schemes, or creating new ones.¹

There is a growing realization, however, that without the right incentives on the supply side of the health system, insurance coverage expansions may not result in more people getting the care they need. And the number of people suffering financial hardship as a result of using health services may not be reduced. In some settings, budgets, salaries and capitation payments leave providers with little financial incentive to deliver extra care, and in such settings pay-for-performance (P4P) is often being introduced in an effort to increase the quantity and quality of care delivered. In other settings, fee-for-service (FFS) is believed to have led to over-provision of care, inappropriate care, and unnecessarily expensive care. In such settings, insurance coverage expansion may lead to people receiving unnecessary care and suffering financial hardship in the process.

While China is not the only emerging economy where such concerns have been investigated (there is evidence also for Indonesia (cf. Sparrow et al. 2013) and the Philippines (cf. Gertler and Solon 2002)), it is China where the evidence is strongest and where the strongest case has been made that without reform to the country's provider

¹ The World Bank has recently published 22 case studies of such initiatives:
<http://www.worldbank.org/en/topic/health/publication/universal-health-coverage-study-series>.

payment system, the government's push toward UHC will not achieve its objectives (cf. e.g. Wagstaff et al. 2009b; Yip et al. 2010; Yip et al. 2012).

We report in this paper the results of a prospective provider payment reform experiment in two counties in Shandong province. Our study estimates – at the level of the township health center (THC) – the relative effects on the cost and quality of care in relation to drug prescribing of a shift from fee-for-service (FFS) to two alternatives: a capitated global budget (CGB); and a combination of CGB and pay-for-performance (P4P). The payment reform was a part of the province's official response to the central government's encouragement of payment reform pilots. As such, political factors inevitably influenced the pilots, and while the pilots in these two counties were planned as a prospective randomized control trial, in some respects the implementation deviated from the original design; we discuss – and where possible try to quantify – the effects of these deviations, and employ the best available methods to avoid any resultant biases. Our experiment sheds light on the ability of supply-side reforms like a shift to P4P to nudge a health system toward the broad goal of UHC: ensuring everyone gets care that is both appropriate to their needs and affordable.

While the focus is on China, our study contributes to a small but growing literature on the effects of P4P initiatives. As we report in the literature review section below, a recent Cochrane Review (Witter et al. 2012) concluded that while P4P is an increasingly popular payment method globally, far too few studies to date have been based on rigorous designs to be able to draw general conclusions. Our experiment also offers some interesting angles compared to other P4P pilots: the starting point is FFS and the goal is to reduce unnecessary care, in contrast to the more common starting point which is a budget system where the goal is to incentivize more care; and the payment scheme we introduce is a

penalty (or ‘stick’) scheme rather than a reward (or ‘carrot’) scheme that is more commonly encountered.

Literature review

A 2012 Cochrane Collaboration systematic review of P4P in low- and middle-income countries (Witter et al. 2012) concluded: “the current evidence base is too weak to draw general conclusions; more robust and also comprehensive studies are needed... Its effects depend on the interaction of several variables, including the design of the intervention (e.g. who receives payments, the magnitude of the incentives, the targets and how they are measured), the amount of additional funding, [and] other ancillary components...”

One of the two developing-country P4P studies that has used a rigorous design was conducted in the Philippines: it found improvements in the quality of care as measured using clinical vignettes (Peabody et al. 2011). The other study was conducted in Rwanda: it found that P4P raised antenatal quality, the number of child preventive care visits, and facility-based deliveries, but had no effect on antenatal care visits or immunization rates (Basinga et al. 2011).

Study setting

Starting in 2003, China has taken major steps to expand insurance coverage, initially to the rural population and more recently to the informal sector (including migrant workers) in cities (Wagstaff et al. 2009b; Wagstaff et al. 2009c; Yip et al. 2012). This coverage expansion has encouraged use of services (one of the policy goals), but – at least in the early years of coverage expansion – did not reduce the amount people pay out of pocket (the other ‘financial protection’ goal) (Wagstaff et al. 2009a; Babiarz et al. 2010; Hou et al.

2013). One explanation (cf. e.g. Wagstaff et al. 2009b) is that because they have traditionally been paid FFS, providers face a financial incentive to deliver more care – especially drugs and tests whose prices allow a profit markup – once people acquire insurance coverage. There is, in fact, evidence consistent with this hypothesis (Liu and Mills 1999; Lu 2014). Inappropriate prescribing behavior by physicians in village clinics and township health centers (THCs) is now seen by some (see e.g. Ministry of Health Center for Health Statistics and Information 2009) to be a significant challenge facing the country’s primary care system. Of particular concern is the overuse of intravenous injection and antibiotics, the consequences of which include not just a heavier financial burden on families, but also damaging effects on patient health, including adverse drug reactions, drug resistance, protracted illness, and even death (Le Grand et al. 1999; Edwards and Aronson 2000; Hu et al. 2001; Dong et al. 2008).

It was hoped that an essential drugs policy would tackle the problem of inappropriate drug prescribing. However, while recent research found evidence of negative effects on average outpatient expenditure per visit, no effects were found on irrational drug use, including appropriate use of antibiotics (Yang et al. 2012; Luo et al. 2013; Sun et al. 2012). Observers argued that what is needed as well was a wholesale reform to the country’s provider payment system (cf. e.g. Wagstaff et al. 2009b; Yip et al. 2010; Yip et al. 2012). In the early 2000s, some local rural health insurance schemes (known as the New Cooperative Medical Scheme or NCMS) began exploring alternative ways to pay THCs and village clinics (Meng 2005); this reform process was strongly encouraged by the central government in the late 2000s. These reforms have not, however, been evaluated with rigorous designs, and there is high chance of selection bias in those that have (Wang et al. 2011; Cheng 2013).

Research design and intervention

Study sites

This study was conducted in 2011-2012 in two counties of Shandong province, located in the country's eastern coastal area and one of China's more economically developed regions (a per capita GDP in 2012 of CNY 51,897, about \$US 8,257). The two study counties, Junan and Liangshan, are located close to one another in the west and southwest parts of the province, and have similar economic conditions. These counties were selected because they have had a long-standing relationship with Shandong University and have been used as experimental sites in the past by the provincial government.

Participating facilities and study design

The experiment covers all the THC's and the village clinics supervised by the THC's in the two counties (data were collected from just two village clinics per THC; see below). Figure 1 shows the study design, and Figure 2 the study timeline.

Prior to July 1, 2011, all THC's and village clinics in the two counties were paid FFS by the payer – the new cooperative medical scheme or NCMS.(Wagstaff et al. 2009a) After a series of workshops on the upcoming payment reform held for NCMS officials and THC managers, the payer switched to one of two new payment methods for all THC's for outpatient services.

Half the THC's (Group A) were switched to a mix of capitation and P4P; the other half (Group B) were switched essentially to capitation. Details of the two payment schemes and the assignment mechanism are presented below.

For both groups, the capitation budget was calculated the same way, covering outpatient services provided by the THC and all the village clinics under the THC's supervision. During the study period, the NCMS sets aside 30% of its budget for outpatient reimbursement. This was then divided by the population size to arrive at a capitation rate. Each THC's capitation budget is thus the capitation rate times the total population in its town.

It had also been envisaged that there would be a third group of townships where FFS would continue to be the payment method for both THCs and village clinics – a genuine control group. However, because of the aforementioned policy shift of the central government, officials at local level were keen to see all townships shift immediately from FFS. This meant that the study is not able to estimate the effects of moving away from FFS, but can estimate only the extra effects associated with combining capitation with P4P rather than shifting to capitation alone.

Payment scheme for Group A: The capitation-plus-P4P group

THCs in Group A (the capitation-plus-P4P group) received 80 percent of their CGB at the start of each quarter. During the first month of the next quarter, an assessment was undertaken by the study team of the quality of care delivered by the THC and the village clinics it supervises; the team then sent a report to the payer. About four weeks later, the THC received a part of the previous quarter's budget that had been withheld, the fraction depending in a simple and well-publicized way on its performance on the quality assessment. The P4P schedule entailed penalties but no bonuses: THCs lost money from their 20 percent withhold whenever a monitored item of care (delivered by them or the village clinics they supervise) deviated from the quality standards agreed beforehand. The

theoretical maximum that a THC in Group A could get was its full 20 percent, but none did so.

The study design did not specify how the THC (the budget holder) pays its village clinics. However, THCs have the authority to pay performance-related bonuses, and of course have an incentive to improve the performance of the village clinics they supervise, either through exhortation or a formal or informal bonus system.

Table 1 shows the quality indicators used, their definition, the targets and penalties associated with shortfalls or overshoots relative to the target, the penalty points involved, the frequency with which data were collected, and the quarters in which each indicator was used to compute the P4P payment. Prescription quality evaluation standards were based on international guidelines and a prescription survey at the end of 2009; they also incorporated feedback from local THC and village clinic doctors. Monitoring data used to calculate penalties were collected from the THC and two randomly selected village clinics supervised by the THC (on average a THC supervises around 36 village clinics). Some indicators (e.g. diagnostic skills) were collected during a quarterly assessment only on a quarterly basis (and not necessarily every quarter as indicated in Table 1). Other data (e.g. prescriptions using steroids) were collected on a monthly basis. At the end of each of the first two months of each quarter, the NCMS office randomly selected two days from the calendar during the previous month, and assessed all the prescriptions of the THC and the two sampled village clinics written on those two days. In the third month of each quarter, the research team selected at random one day in each of the previous three months, and assessed all the prescriptions of the THC and two sampled village clinics. In the first two months of each quarter therefore there were two days' worth of prescriptions assessed by the NCMS and one day's worth assessed by the study team: a weighted average of the three

scores was computed with a weight of 0.1 on each of the two scores from the NCMS assessment and a weight of 0.8 on the study team's assessment. In the fourth quarter, there was just one score (from the NCMS assessment) and this was given a weight of 1.0. The quality of the two selected village clinics within each township was considered to represent the quality of all village clinics within the township, so the total deducted points for the two village clinics combined were added to the points deducted from the THC to arrive at a grand total for the township. Each one-point deduction in this grand total resulted in a deduction from the township's allocation of RMB 100 (US\$ 15-16 at the time).

Payment scheme for Group B: The capitation group

THCs in Group B were paid essentially by capitation. The intention was, in fact, that they would be paid *entirely* by capitation, and the differential change between Group A and Group B would show the additional effect of P4P over and above a move from FFS to GCB. In the event, local officials were uncomfortable about paying Group B THCs the full 100 percent of the GCB in advance. Recall that previously the payer had paid THCs using FFS, and had thus paid for services that had actually been delivered; a shift from FFS to pure capitation involved an uncomfortably large leap of faith in the service providers on the part of the payer.

So it was decided to withhold 20 percent of the capitation-based budget of Group B THCs pending an assessment at the end of the year of what they had actually delivered. Unlike Group A facilities, however, Group B facilities had to wait until the start of the next year before they received their share of the withheld 20 percent. And, in contrast to Group A facilities who knew the rules in advance, Group B facilities did not; they were not told on what basis the authorities would decide how much of the withheld 20 percent the facility

would receive, only that the authorities would undertake “a comprehensive assessment”. In the event, the authorities decided at the end of year not to assess the quality of care delivered by Group B facilities. Of course, during the period of the experiment, Group B THC’s did not know that the comprehensive assessment would turn out to be so light, and it seems reasonable to assume that they too will have altered their behavior; our estimates will therefore be underestimates of the true effects of the P4P intervention.

Randomization

The original research design envisaged that THC’s would be assigned to groups A and B through a matched-pair randomization procedure (King et al. 2009). Within each county, townships were to be ranked by their per capita income, and formed into adjacent pairs; starting with the poorest pair of townships, via a toss of a coin, one of each pair would be randomly assigned to Group A and the other would go to Group B. In the event, due apparently to a misunderstanding, a somewhat different randomization approach was used: the townships were ordered by per capita income; one was selected at random and assigned to Group A; the next richest was assigned to Group B, the next richest to Group A, and so on until all townships richer than the randomly selected township had been assigned; the remaining townships were then assigned using the same approach, starting with the poorest which was assigned to the opposite group from the one the richest township had been assigned to. The authorities in one county (Liangshan) then did some further reshuffling, citing concerns about geographic clustering. We explore below the effects on baseline balancing of these two departures from the research protocol, and employ statistical methods that try to eliminate any biases that they may have induced.

Outcomes

We focus on prescription-level outcomes that we were able to assess independently from the monitoring system. Data on all our outcome indicators come from random samples of prescriptions, collected from each THC and village clinic during the baseline and endline surveys; 50 prescriptions were selected from each THC; 20 from each selected village clinic. Prescriptions collected covered the month of April during the year of the survey in 2011 and 2012.

All our outcomes were incentivized by the P4P scheme: whether the prescription contained 6+ drugs; whether the prescription contained 2+ antibiotics; whether the prescription involved intravenous injection; and whether the prescription contained steroids. We also looked at the effects of the P4P scheme on the cost of the consultation, which – while not directly incentivized – is clearly an outcome of interest, especially in the Chinese context. Most visit records in China do not itemize separately the consultation fee and the charge for the medicines. However, the former has historically been small, and in our data for those cases where both numbers were recorded the difference was indeed small. (The cost is recorded on the prescription because even though China’s rural residents are now mostly covered by the NCMS, some are not, and those who are covered pay a share of the cost.)

Statistical analysis

We estimate the impacts of the P4P scheme using a multivariate specification of the difference-in-difference model, running separate regressions for THCs and village clinics. Our approach is thus similar to that of the authors of a P4P experiment in Rwanda (Basinga et al. 2011) whose initial random assignment of facilities was also compromised by

political factors. We regress the outcome variable for a facility on a dummy variable indicating whether the facility was paid under the GCB+P4P scheme that year, a facility fixed effect, a year indicator, and the gender and age (in bands) of the patient. Unlike a simple comparison of the two groups' means in the endline data, which would be inappropriate in this case given that the assignment between the two groups did not follow the original pairwise randomization protocol, the fixed-effect model with covariates nets out any confounding effects of differential changes in the demographic mix of patients between the two groups, and controls for the effects of unobserved (and observed) time-invariant facility characteristics. We calculated robust standard errors allowing for clustering at the township level, the year level, and the intersection of the two (township-by-year) (Cameron et al. 2011).

On the village clinic sample we find good balancing on observables at baseline, but for the THC sample we did not. To further reduce any biases due to these baseline differences, we used coarsened exact matching (CEM) (King and et al. 2007; King et al. 2010) to come up with a subsample of THCs that are comparable at baseline, discarding both treatment and control THCs that contribute to a lack of balance at baseline, and then running the fixed-effects regression model on the CEM-trimmed subsample. CEM involves deciding on a list of matching variables, converting any continuous variables into categorical variables, and then matching observations (in this case across the treatment and control groups) according to the categories they fall into on the matching variables. A researcher can either let CEM determine the cutpoints or the number of categories, or can specify them. We included in the list of matching variables (at THC level) the means of the outcomes in Table 2 along with the means of the corresponding binary variables for the prescription variables indicating whether or not a prescription had the characteristic in

question (e.g. involved steroids). We opted to have each matching variable coarsened into five categories (recall all variables are means measured at the THC level).

Results

Baseline data

Baseline values for the full sample are shown in Table 2. The numbers highlight the challenges that China faces in curbing over-prescription of drugs. Around 14 percent of prescriptions contain six or more drugs, around 12 percent of prescriptions contain multiple antibiotics, 31 percent of prescriptions involve intravenous injection, and 12 percent of prescriptions contain steroids.

In the village clinic data, none of the baseline differences between groups A and B are statistically significant at the five or ten percent levels. In the THC data, by contrast, three of the five outcome variables differ significantly at baseline at the five percent level; the treatment THCs have higher initial values of all outcome variables, often dramatically higher. The third block of data in Table 2 shows how the balancing for the THC data looked in the original assignment – after the research team had done the randomization (albeit not the originally envisaged randomization method) but before the local government officials in Liangshan county reshuffled some townships between the treatment and control groups: interestingly, no significant differences in age, gender and the five outcomes emerge. The fourth block of data in Table 2 shows how the balancing would have looked if the original matched-pair randomization procedure had been adhered to and there had been no interference by local government officials in the assignment to the two groups: here, unsurprisingly, no significant differences emerge. Figure 3 shows the average absolute standardized bias in these three assignments: the average bias in the actual assignment is

over twice that in the original assignment; had the original matched-pair randomization procedure been used, the bias would have been just 30 percent of the bias in the actual assignment. Of course it was the actual assignment that was implemented, so the balancing results for the original assignment and matched-pair random assignment are academic. What we *can* do, however, is focus on a subset of the THCs in the two actual assignment groups, using CEM to choose the subsets that minimize baseline differences. The last block of data in Table 2 shows the balancing for this subsample. The baseline values are still somewhat higher in the (trimmed) treatment group than in (trimmed) control group but none of the differences is significant at the five percent level; moreover, as Figure 3 shows, the average absolute standardized bias for the CEM-trimmed subsample is 50 percent smaller than that associated with the actual assignment for the full sample. (The bias is, of course, larger than it would have been if the original assignment had been kept or if the matched-pair randomization protocol had been implemented.)

Table 3 shows the THC baseline differences by county. Two points emerge. First, Junan has lower values of the quality indicators than Liangshan. Indeed, by the time the intervention was implemented, Junan's quality indicators were already close to or below the thresholds used to compute the quality-related penalties in the P4P scheme. We would as a result expect the P4P scheme to have had less impact (if any) in Junan – a hypothesis we test below. Second, the balancing is better at baseline in Junan (where there was no political interference in the assignment process) than in Liangshan (where there was). Indeed, in Junan we see just one statistically significant difference at baseline, whereas in Liangshan we see significant differences at the five percent level in three of the outcomes. The CEM routine actually retains all the Junan townships and weights them equally, and all the trimming seen in Table 2 was done in the Liangshan THC sample. In Table 3, the

CEM-trimmed Liangshan subsample emerges with smaller and mostly insignificant differences between the treatment and control THCs.

Penalties

Table 4 shows the average penalty points per township for each of the first three quarters of the experiment, and for the first three quarters combined. On the outcome indicators used in the study, the average Junan township incurred half as many penalty points as the average Liangshan township, reflecting the fact that Junan was closer to or at the thresholds by the time the experiment began. The higher penalty points incurred by Junan on the skill test are explained at least in part by the fact that Junan had a higher penalty rate (1/5 compared to Liangshan's 1/100). Likewise, the higher penalty points on patient satisfaction are also explained at least in part by the fact that from the second quarter Junan moved to a higher threshold (100 while Liangshan stayed at 80) and to a higher penalty rate (1/2 while Liangshan stayed with 1/20).

Effect estimates

Table 5 shows our estimates of the impacts of P4P obtained using the fixed effects model. The coefficient β shows the additional effect of moving from FFS to CGB+P4P over and above moving from FFS to CGB. Where the outcome is also a dummy variable, the coefficient gives the percentage point increment associated with P4P; in the case of the cost variable, the coefficient gives the additional effect associated with P4P in Chinese RMB. We also report for both types of variable the percentage effect.

The village clinic results show that P4P has a dampening effect on all but one outcome indicator, but none of the effects is statistically significant at the five percent level (henceforth, statistical significance is always assessed at the five percent level unless stated

otherwise). Somewhat larger dampening effects are seen in the CEM-trimmed THC subsample, but again none of the effects is significant at the five percent level.

Table 6 shows impact estimates by county; in Junan, the results are for the full sample, while in Liangshan they are for the CEM-trimmed subsample. In Junan, we see no significant effects for the THCs, but significant negative effects for two of the five outcomes for the village clinics. In Liangshan, by contrast, for three of the five outcome indicators we see negative effects for THCs that are significant at the five percent level, and a fourth that is significant at the ten percent level. At village level, in Liangshan, we see no significant negative effects, but oddly we do see a significant positive effect on multi-antibiotic prescriptions. There is a perfectly plausible explanation of why we see different effects in the two counties among THCs: as was seen in Table 3, THCs in Junan were already close to or below the thresholds by the time the intervention began (and hence did not stand to gain financially from further reductions), while in Liangshan THCs went into the experiment with quality indicators that exceeded the thresholds (and hence did stand to gain financially). The differences between the THC effects and the village clinic effects are not as stark as we expected, but this may reflect the fact that while the local governments did not alter the FFS payment system for village clinics, the payment to THCs did reflect the performance of village clinics, and some THCs may therefore have explored ways to incentivize better performance by their reporting clinics.

Discussion

Like many ‘field’ experiments involving government ‘actors’, this experiment was not as ‘clean’ as the study team would have liked. Because of political pressures from central government to shift away from FFS, local government officials were unwilling to retain the

original ‘pure’ control group in which facilities would continue to be paid FFS. The study team thus had to settle for an experiment that compared a shift from FFS to pure capitation with a shift from FFS to capitation combined with a performance bonus. Even here, the experiment was subject to political ‘interference’: being used to a payment system that paid by results (FFS), local government officials balked at the idea of handing out 100 percent of a facility’s capitation payment in advance without any mechanism to assure delivery of services; they decided instead to withhold 20 percent on the understanding that this would be paid at the end of the year conditional on satisfactory performance. However, it was not made clear until after the experiment, how performance would be assessed, and there is a possibility that the control group may have been expecting a quality audit at the end of the year, and hence may have altered its behavior vis-à-vis prescribing quality; the effect would be to bias our results downwards.

In the randomization process, too, messiness crept in. The randomization process in both counties deviated from the matched-pair procedure originally envisaged; while this raised the average absolute standardized bias above what it would have been the case if the protocol had been followed, it did not do so enough to cause any significant differences at baseline between the treatment and control groups. Much more damaging was the decision by local government officials in one of the two counties to swap some of the townships between the treatment and control groups. This led to a considerable increase in average bias at baseline between the two groups, and to significant differences at baseline in three of the five outcome variables. We were forced to resort to trimming (i.e. dropping units from) the treatment and control groups using coarsened exact matching in an effort to reduce bias (we did so by 50 percent) and to eliminate significant differences at baseline between the treatment and control groups (we succeeded).

The pooled results for the two counties show no effects of P4P at either the village or THC level on cost per visit or on the appropriateness of prescribing behavior, as measured by four incentivized quality indicators capturing whether a prescription contains six or more drugs, whether it contains two or more antibiotics, whether it entailed an intravenous injection, and whether steroids were prescribed. By contrast, the disaggregated results by county reveal quantitatively large and statistically significant negative effects on incentivized quality indicators among THCs in one county (Liangshan) but not in the other (Junan). This is likely to be due to the fact that by the time the experiment began, Junan's THC incentivized quality indicators had fallen below the thresholds that triggered payment penalties while in Liangshan this was not the case; the THCs there thus had an incentive to continue to improve their prescribing quality indicators. An accidental feature of its design thus allowed the experiment to demonstrate an important lesson for the design of P4P schemes involving penalties for poor performance: if thresholds are used, they must 'bite'. The alternative, especially if – as seems inevitable – there is some uncertainty as to where the facilities will be in terms of the threshold by the time the incentive scheme starts, is simply to avoid thresholds altogether, and incentivize marginal improvements in performance whatever the level of performance.

Two further aspects of the results are worthy of discussion. First, even in Liangshan where significant effects of P4P on prescribing behavior are seen, no dampening effects were seen on the out-of-pocket spending. This suggests some compensatory behavior on the part of THC doctors, and that the incentive scheme did not do enough to discourage such behavior. We are not able to establish with our data what this compensatory behavior might have entailed, and an important lesson for future such studies is to ensure that sufficiently detailed data are collected that allows such compensatory behavior to be

tracked. Second, even among the Liangshan THCs where the P4P scheme did reduce inappropriate prescribing, the endline levels are still quite high: most notably, even after the P4P intervention, almost 30 percent of prescriptions in the trimmed THC sample involved an intravenous injection. This suggests that while P4P may be a useful weapon in the war against financially-motivated inappropriate prescribing, it would be worth exploring complementary interventions, such as publicizing individual facilities' rates of inappropriate prescribing in the hope of shaming them into changing their behavior even faster.

Table 1: Quality assessment tool used for P4P in Group A

Quality indicator	Definition	Target and penalty	Frequency of assessment	Quarter			
				1	2	3	4
Qualified prescriptions	A 'qualified' written outpatient prescription should include the time and date, the patient's name, their gender and age, their address, the diagnosis, the drug type, the drug dosage, ingestion method, and the signatures of the doctor and the pharmacist.	The target is 100 percent. For each percentage point below 100, half a point was deducted.	Quarterly	✓	✓	✓	✓
Prescriptions containing 6 or more drugs		The proportion should be below 10 percent. For each percentage point in excess of 10, half a point was deducted.	Monthly	✓	✓	✓	✓
Divided prescriptions	A 'divided' prescription is one where the same patient is given more than one prescription for the same diagnosis on the same day in the same facility.	The target is zero. Each occurrence results in a deduction of 1/10 of a point.	Monthly	✓	✓	✓	✓
Prescriptions combined multiple antibiotics		For THCs the target was 10 percent in Junan, and 15 percent in Liangshan. For village clinics, the target was 15 percent in Junan, and 20 percent in Liangshan. For each percentage point in excess of the target, half a point was deducted.	Monthly	✓	✓	✓	✓
Prescriptions using intravenous transfusion		For THCs the target was 25 percent in both counties; for village clinics, the target was 35 percent in both counties. For each percentage point in excess of the target, RMB 50 were deducted.	Monthly	✓	✓	✓	✓
Prescriptions using steroids		For THCs the target was 10 percent in both counties; for village clinics, the target was 22 percent in Junan, and 20 percent in Liangshan. For each percentage point in excess of the target, half a point was deducted.	Monthly	✓	✓	✓	✓
Score on test of skill in diagnosing and treating chronic illnesses	A bank of vignette-style questions was designed following authoritative treatment guidelines. Doctors in village clinics and a subset of doctors in a THC are asked to answer ten randomly selected questions from the question bank.	The target is 80 marks. In Junan for each mark below 80, 1/5 of a point was deducted. In Liangshan, each point below 80 resulted in a 1/100 of a point deduction.	Quarterly	✓	✓	✓	
Patient satisfaction	On exiting the facility, randomly selected patients are asked to indicate their satisfaction on a three-point scale on six questions covering the service attitudes, technical skills and communication of the medical staff, the amount spent, and the physical environment of	The target score was 80 percent in both counties initially; Junan raised it to 100 from the second quarterly evaluation onwards. In Junan, initially each percentage point below the 80 point target resulted in a 1/20 of a point deduction; from the second quarter in Junan each	Quarterly	✓	✓	✓	

	the facility. Patient satisfaction is computed as the sum of patients who are very satisfied or satisfied, divided by the number of interviewed patients times six.	point below the 100 target resulted in a 1/2 point deduction. In Liangshan, each point below the 80 point target resulted in a 1/20 of a point was deducted.					
Environment and hygiene of the village clinic	The clinic should be sanitary and hygienic, and everything should be appropriately stored.	If according to the subjective evaluation of the assessor the facility is not satisfactory, 1/10 of a point was deducted.	Every half year		✓		
Environment and hygiene of the THC	The THC should be sanitary and hygienic, and everything should be appropriately stored.	If according to the subjective evaluation of the assessor the facility is not satisfactory, 1/10 of a point was deducted.	Every half year		✓		

Notes: In the case of data collected every month, the NCMS office collected the data in the first two months of each quarter, and the NCMS office and the research team together collected the data for the third month and collated the data for the entire quarter. The fourth quarter was an exception, when the NCMS office collected the data by itself. In the case of the case of the data collected on a quarterly basis, the data were collected only at the end of the quarter. Data for the last four indicators listed were not collected in the fourth quarter, and data for the last two were not collected in the first and third quarters either.

Table 2: Baseline differences

	Group A (CGB+P4P)			Group B (CGB)			Difference	
	N	Mean	SD	N	Mean	SD	Mean*	p-value ^{&}
<i>Village</i>								
Age in years	1032	39.517	23.404	976	41.077	22.347	-1.559	0.629
Male	1142	0.651	0.477	1059	0.687	0.464	-0.036	0.345
Cost	906	29.592	39.572	986	27.201	39.649	2.392	0.633
Prescription contained 6 or more drugs	1163	0.130	0.336	1115	0.073	0.260	0.057	0.103
Prescription contained 2+ antibiotics	1163	0.114	0.317	1114	0.151	0.358	-0.037	0.368
Prescription involved intravenous injection	1163	0.357	0.479	1115	0.331	0.471	0.026	0.786
Prescription contained steroids	1163	0.103	0.304	1115	0.104	0.305	-0.001	0.980
<i>THC – actual</i>								
Age in years	725	38.989	24.663	568	42.284	23.200	-3.295	0.322
Male	765	0.512	0.500	607	0.478	0.500	0.035	0.416
Cost	581	43.759	80.272	514	40.179	64.161	3.580	0.716
Prescription contained 6 or more drugs	780	0.240	0.427	631	0.116	0.320	0.124	0.024
Prescription contained 2+ antibiotics	780	0.149	0.356	630	0.071	0.258	0.077	0.018
Prescription involved intravenous injection	780	0.356	0.479	631	0.214	0.410	0.143	0.104
Prescription contained steroids	780	0.172	0.377	631	0.098	0.298	0.074	0.031
<i>THC – original assignment[@]</i>								
Age in years	665	37.802	24.456	628	43.226	23.369	-5.425	0.100
Male	709	0.505	0.500	663	0.489	0.500	0.016	0.693
Cost	525	44.524	83.444	570	39.826	62.153	4.698	0.645
Prescription contained 6 or more drugs	721	0.204	0.403	690	0.164	0.370	0.040	0.505
Prescription contained 2+ antibiotics	721	0.135	0.342	689	0.093	0.291	0.042	0.250
Prescription involved intravenous injection	721	0.304	0.460	690	0.281	0.450	0.023	0.800
Prescription contained steroids	721	0.154	0.361	690	0.123	0.329	0.031	0.414
<i>THC – block randomization (hypothetical) #</i>								
Age in years	628	42.417	23.730	665	38.565	24.272	3.852	0.252
Male	665	0.508	0.500	707	0.487	0.500	0.022	0.595
Cost	564	35.229	52.100	531	49.354	89.757	-14.126	0.151
Prescription contained 6 or more drugs	685	0.191	0.394	726	0.178	0.383	0.014	0.822
Prescription contained 2+ antibiotics	685	0.096	0.295	725	0.131	0.338	-0.035	0.339
Prescription involved intravenous injection	685	0.302	0.460	726	0.284	0.451	0.018	0.835
Prescription contained steroids	685	0.136	0.343	726	0.142	0.349	-0.006	0.873
<i>THC – CEM[§]</i>								
Age in years	564	39.660	24.055	568	42.284	23.200	-2.624	0.409
Male	601	0.498	0.500	607	0.478	0.500	0.020	0.654
Cost	484	45.465	87.091	514	40.179	64.161	5.286	0.632
Prescription contained 6 or more drugs	615	0.172	0.378	631	0.116	0.320	0.057	0.179
Prescription contained 2+ antibiotics	615	0.101	0.301	630	0.071	0.258	0.029	0.081
Prescription involved intravenous injection	615	0.276	0.448	631	0.214	0.410	0.063	0.425
Prescription contained steroids	615	0.129	0.335	631	0.098	0.298	0.030	0.241

* The mean of Group A – the mean of Group B. [&] The p-value is for a simple test of difference between two means; its value allows for clustering at the county level. [@] A simulated assignment corresponding to the original research protocol: townships were ranked in ascending order of per capita income in each county, for each township a random number was generated (using the same seed value throughout), and then the township in each pair with the larger random number was assigned to Group A. [§] CEM (coarsened exact matching) undertaken using in addition to the variables listed above the actual numbers of drugs prescribed, antibiotics, injection drugs, and steroids; each was coarsened into five categories.

Table 3: THC baseline differences by county

	Group A (CGB+P4P)			Group B (CGB)			Difference	
	N	Mean	SD	N	Mean	SD	Mean*	p-value&
<i>Junan</i>								
Age in years	419	37.084	23.655	365	41.630	23.667	-4.547	0.221
Male	450	0.527	0.500	400	0.505	0.501	0.022	0.663
Cost	436	48.564	90.403	408	45.628	69.433	2.936	0.800
Prescription contained 6 or more drugs	458	0.170	0.376	422	0.159	0.366	0.012	0.806
Prescription contained 2+ antibiotics	458	0.098	0.298	421	0.057	0.232	0.041	0.027
Prescription involved intravenous injection	458	0.238	0.426	422	0.263	0.441	-0.025	0.808
Prescription contained steroids	458	0.127	0.333	422	0.109	0.312	0.018	0.484
<i>Liangshan (before CEM)</i>								
Age in years	306	41.598	25.791	203	43.458	22.347	-1.860	0.761
Male	315	0.492	0.501	207	0.425	0.496	0.067	0.373
Cost	145	29.309	31.386	106	19.204	29.433	10.105	0.253
Prescription contained 6 or more drugs	322	0.339	0.474	209	0.029	0.167	0.310	0.004
Prescription contained 2+ antibiotics	322	0.221	0.415	209	0.101	0.301	0.120	0.077
Prescription involved intravenous injection	322	0.525	0.500	209	0.115	0.320	0.410	0.001
Prescription contained steroids	322	0.236	0.425	209	0.077	0.267	0.160	0.025
<i>Liangshan (after CEM) §</i>								
Age in years	145	47.103	23.725	203	43.458	22.347	3.645	0.571
Male	151	0.411	0.494	207	0.425	0.496	-0.015	0.844
Cost	48	17.309	37.513	106	19.204	29.433	-1.896	0.759
Prescription contained 6 or more drugs	157	0.178	0.384	209	0.029	0.167	0.150	0.067
Prescription contained 2+ antibiotics	157	0.108	0.312	209	0.101	0.301	0.008	0.815
Prescription involved intravenous injection	157	0.389	0.489	209	0.115	0.320	0.274	0.005
Prescription contained steroids	157	0.134	0.342	209	0.077	0.267	0.057	0.461

* The mean of Group A – the mean of Group B. & The p-value is for a simple test of difference between two means; its value allows for clustering at the county level. # A simulated assignment corresponding to the original research protocol: townships were ranked in ascending order of per capita income in each county, for each township a random number was generated (using the same seed value throughout), and then the township in each pair with the larger random number was assigned to Group A. § CEM (coarsened exact matching) undertaken using in addition to the variables listed above the actual numbers of drugs prescribed, antibiotics, injection drugs, and steroids; each was coarsened into five categories.

Table 4: Penalty points incurred, by county

	Liangshan				Junan			
	Q1	Q2	Q3	Average	Q1	Q2	Q3	Average
Qualified prescriptions	33.1		2.9	18.0	22.2	10.8	2.7	11.9
Prescriptions containing 6 or more drugs	6.8	1.0	0.5	2.8	1.0	1.8	0.0	0.9
Divided prescriptions	0.0	0.0	0.2	0.1	0.3	0.0	0.0	0.1
Prescriptions combined multiple antibiotics	4.2	1.9	1.9	2.7	0.1	1.5	0.2	0.6
Prescriptions using intravenous transfusion	5.6	8.3	5.5	6.5	3.6	2.0	2.3	2.6
Prescriptions using steroids	7.4	3.5	2.0	4.3	6.2	1.3	4.5	4.0
Score on test of skill in diagnosing and treating chronic illnesses	0.2	0.5	0.3	0.3	16.7	4.4	12.3	11.1
Patient satisfaction	0.1	0.0	0.0	0.0	0.2	7.6	4.9	4.2
<i>Average across all indicators</i>	<i>7.2</i>	<i>2.2</i>	<i>1.6</i>	<i>4.3</i>	<i>6.3</i>	<i>3.7</i>	<i>3.4</i>	<i>4.4</i>
<i>Average across outcome indicators in the study</i>	<i>6.0</i>	<i>3.7</i>	<i>2.5</i>	<i>4.0</i>	<i>2.7</i>	<i>1.7</i>	<i>1.8</i>	<i>2.0</i>

Notes: Q1 means first quarter, and so on. Environment and hygiene indicators are omitted from the table, as they did not, in the event, attract any penalty points. Qualified prescriptions were not assessed in Liangshan in Q2.

Table 5: Impact estimates – main results

	Fixed effects model results					Endline values, P4P group	
	N [§]	β	95% CI	p-value	β (%)*	N [#]	Mean [#]
<i>Village</i>							
Cost	3273	-3.132	(-10.80 - 4.53)	0.423	-11%	804	18.335
Prescription contained 6 or more drugs	3876	-0.024	(-0.07 - 0.02)	0.312	-19%	1011	0.101
Prescription contained 2+ antibiotics	3873	0.054	(-0.02 - 0.13)	0.169	48%	1010	0.148
Prescription involved intravenous injection	3874	-0.063	(-0.16 - 0.03)	0.176	-18%	1010	0.272
Prescription contained steroids	3874	-0.033	(-0.09 - 0.03)	0.278	-32%	1010	0.100
<i>THC (on CEM sub-sample)</i>							
Cost	1724	-12.265	(-29.04 - 4.51)	0.152	-27%	367	31.586
Prescription contained 6 or more drugs	2108	-0.038	(-0.10 - 0.02)	0.202	-22%	486	0.181
Prescription contained 2+ antibiotics	2108	-0.036	(-0.09 - 0.01)	0.166	-35%	486	0.113
Prescription involved intravenous injection	2108	-0.031	(-0.13 - 0.07)	0.552	-11%	486	0.344
Prescription contained steroids	2108	-0.051	(-0.12 - 0.02)	0.175	-40%	486	0.134

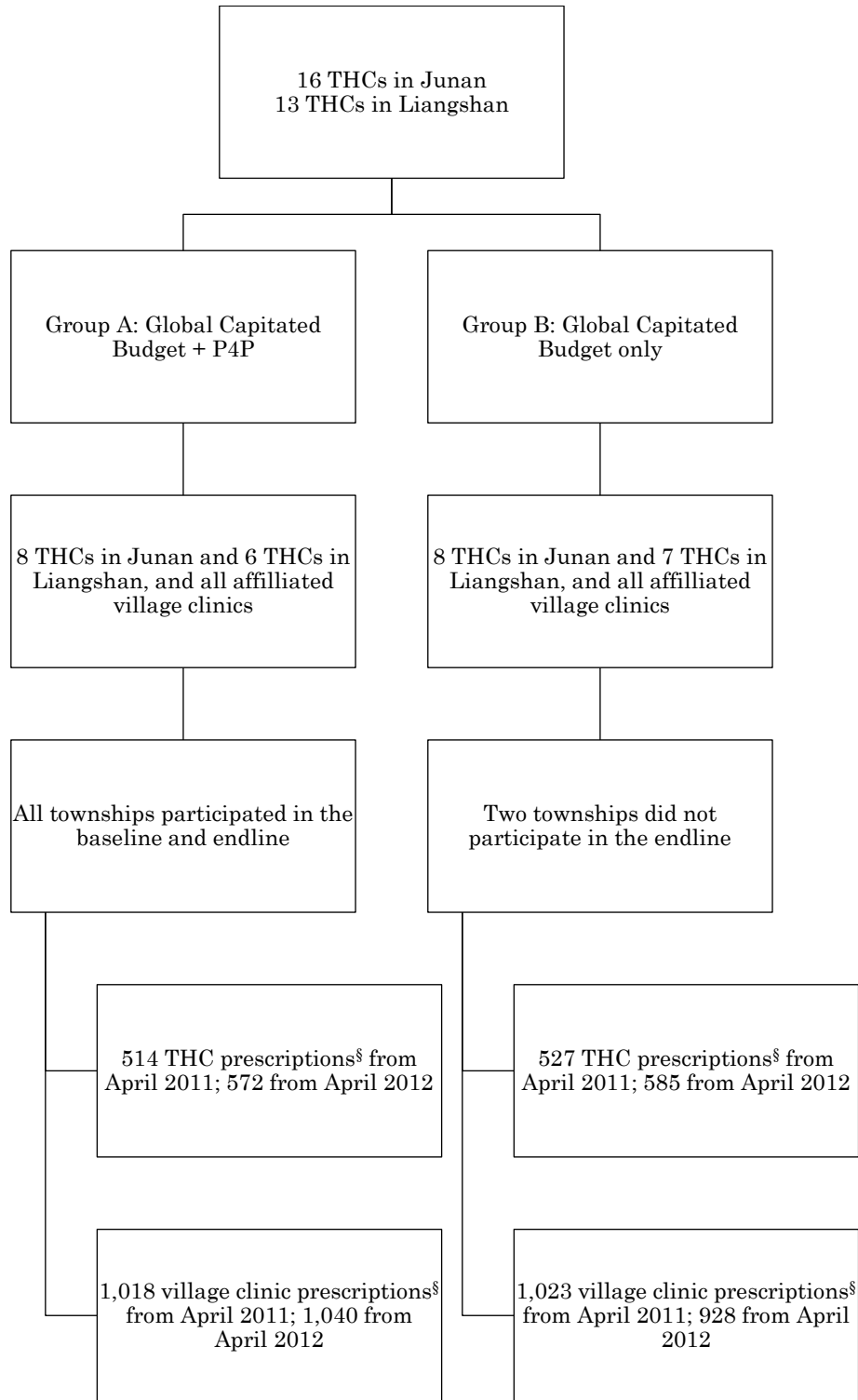
[§]Baseline and endline, and both groups. *Equal to $(\beta/\text{baseline mean}) \times 100$ where the baseline mean is for the P4P group. [#]Endline, P4P group only.

Table 6: Impact estimates – results by county

	Fixed effects model results					Endline values, P4P group	
	N [§]	β	95% CI	p-value	β (%)*	N [#]	Mean [#]
<i>Junan – village</i>							
Cost	2068	-2.733	(-12.60 - 7.14)	0.587	-9%	506	20.673
Prescription contained 6 or more drugs	2201	-0.061	(-0.10 - -0.02)	0.005	-77%	614	0.036
Prescription contained 2+ antibiotics	2199	-0.003	(-0.07 - 0.07)	0.926	-4%	613	0.090
Prescription involved intravenous injection	2199	-0.111	(-0.21 - -0.01)	0.031	-63%	613	0.140
Prescription contained steroids	2199	-0.033	(-0.11 - 0.05)	0.410	-41%	613	0.077
<i>Junan – THC</i>							
Cost	1406	-13.733	(-32.76 - 5.30)	0.157	-28%	326	32.887
Prescription contained 6 or more drugs	1459	0.029	(-0.02 - 0.08)	0.238	17%	350	0.206
Prescription contained 2+ antibiotics	1459	-0.005	(-0.07 - 0.06)	0.873	-5%	350	0.126
Prescription involved intravenous injection	1459	0.104	(-0.01 - 0.22)	0.073	44%	350	0.363
Prescription contained steroids	1459	-0.009	(-0.09 - 0.07)	0.815	-7%	350	0.134
<i>Liangshan – village (after CEM)</i>							
Cost	720	-10.546	(-27.60 - 6.51)	0.226	-31%	111	12.768
Prescription contained 6 or more drugs	981	0.027	(-0.04 - 0.10)	0.459	16%	163	0.166
Prescription contained 2+ antibiotics	981	0.169	(0.00 - 0.34)	0.045	259%	163	0.190
Prescription involved intravenous injection	981	-0.031	(-0.27 - 0.20)	0.797	-6%	163	0.417
Prescription contained steroids	981	-0.064	(-0.19 - 0.06)	0.297	-85%	163	0.092
<i>Liangshan – THC (after CEM)</i>							
Cost	318	3.128	(-15.14 - 21.40)	0.737	18%	41	21.244
Prescription contained 6 or more drugs	649	-0.166	(-0.30 - -0.03)	0.017	-93%	136	0.118
Prescription contained 2+ antibiotics	649	-0.115	(-0.17 - -0.06)	0.000	-106%	136	0.081
Prescription involved intravenous injection	649	-0.316	(-0.38 - -0.25)	0.000	-81%	136	0.294
Prescription contained steroids	649	-0.115	(-0.25 - 0.02)	0.094	-86%	136	0.132

[§]Baseline and endline, and both groups. *Equal to $(\beta/\text{baseline mean}) \times 100$ where the baseline mean is for the P4P group. [#]Endline, P4P group only.

Figure 1: Study design

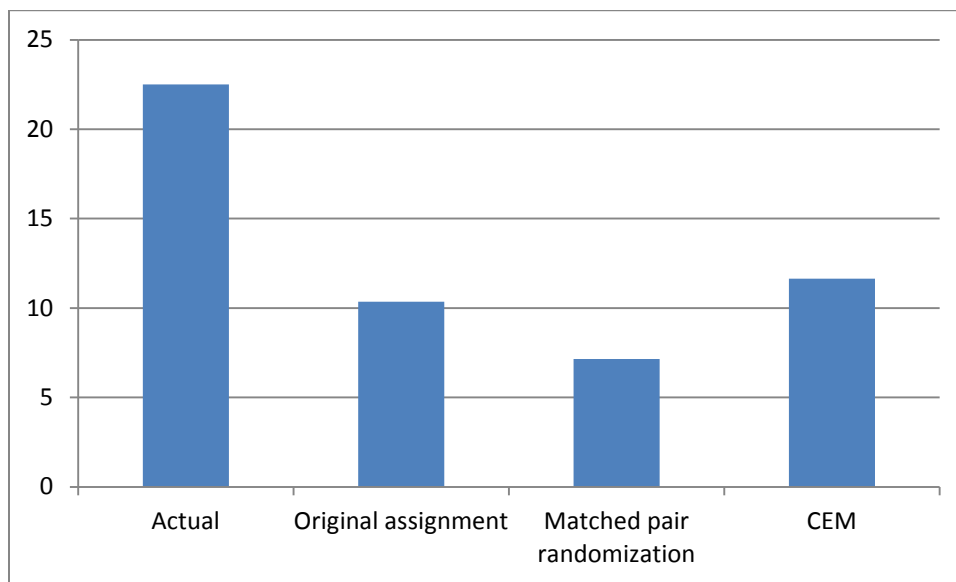


[§]Includes only prescriptions containing complete information on drugs prescribed, the patient's gender and age, and – in the case of the THC – the prescribing department. Not all of these contained the charge for the visit.

Figure 2: Study timeline

	May-June 2011	June 2011	July 1, 2011	July 1 2011-June 30, 2012		July 2012	July-August 2012
Group A: GCB+P4P	Baseline survey	Training workshops on the P4P scheme for local NCMS officials and THC managers	FFS stops. CGB+P4P starts. Facilities receive 80% of GCB at the start of each quarter	THCs and the village clinics for which they are responsible are assessed on quality on a monthly and/or quarterly basis	Four weeks into the next quarter, facilities receive part of the 20% GCB withheld in the previous quarter, depending on their quality score	Part of the 4 th quarter's 20% withhold paid according to 4 th quarter's quality score	Endline survey
Group B: GCB			FFS stops. CGB starts. Facilities receive 80% of GCB at the start of each quarter			Assessment of the previous year. In practice focuses on outpatient costs. The four quarterly 20% withholds are paid in full	

Figure 3: Average absolute standardized bias for THC data



References

- Babiarz, K. S., G. Miller, H. Yi, L. Zhang and S. Rozelle (2010). "New evidence on the impact of China's New Rural Cooperative Medical Scheme and its implications for rural primary healthcare: multivariate difference-in-difference analysis." *BMJ* 341: c5617.
- Basinga, P., P. J. Gertler, A. Binagwaho, A. L. Soucat, J. Sturdy and C. M. Vermeersch (2011). "Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation." *Lancet* 377(9775): 1421-1428.
- Cameron, A. C., J. B. Gelbach and D. L. Miller (2011). "Robust Inference With Multiway Clustering." *Journal of Business & Economic Statistics* 29(2): 238-249.
- Cheng, T. M. (2013). "A pilot project using evidence-based clinical pathways and payment reform in China's rural hospitals shows early success." *Health Aff (Millwood)* 32(5): 963-973.
- Dong, L., H. Yan and D. Wang (2008). "Antibiotic prescribing patterns in village health clinics across 10 provinces of Western China." *J Antimicrob Chemother* 62(2): 410-415.
- Edwards, I. R. and J. K. Aronson (2000). "Adverse drug reactions: definitions, diagnosis, and management." *Lancet* 356(9237): 1255-1259.
- Gertler, P. and O. Solon (2002). "Who benefits from social health insurance? Evidence from the Philippines." *Unpublished Manuscript, University of California, Berkeley and the University of the Philippines.*
- Hou, Z., E. Van de Poel, E. Van Doorslaer, B. Yu and Q. Meng (2013). "Effects of NCMS on access to care and financial protection in China." *Health Econ*: n/a-n/a.
- Hu, S., W. Chen, X. Cheng, K. Chen, H. Zhou and L. Wang (2001). "Pharmaceutical cost-containment policy: experiences in Shanghai, China." *Health Policy Plan* 16 Suppl 2: 4-9.
- King, G., M. Blackwell, S. Iacus and G. Porro (2010). "cem: Coarsened exact matching in Stata." *Stata Journal* 9(4): 524-546.
- King, G. and et al. (2007). "A "Politically Robust" Experimental Design for Public Policy Evaluation, with Application to the Mexican Universal Health Insurance Program." *Journal of Policy Analysis and Management* 26(3): 479-506.
- King, G., E. Gakidou, K. Imai, J. Lakin, R. T. Moore, C. Nall, N. Ravishankar, M. Vargas, M. M. Tellez-Rojo, J. E. Avila, M. H. Avila and H. H. Llamas (2009). "Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme." *Lancet* 373(9673): 1447-1454.
- Le Grand, A., H. V. Hogerzeil and F. M. Haaijer-Ruskamp (1999). "Intervention research in rational use of drugs: a review." *Health Policy Plan* 14(2): 89-102.
- Liu, X. and A. Mills (1999). "Evaluating payment mechanisms: how can we measure unnecessary care?" *Health Policy and Planning* 14(4): 409-413.
- Lu, F. (2014). "Insurance coverage and agency problems in doctor prescriptions: Evidence from a field experiment in China." *Journal of Development Economics* 106(0): 156-167.
- Luo, F., L. Yao, Q. Yao, L. He, H. Liu, Y. Pan, S. Chen, N. Cai and Z. Liu (2013). "Impact of essential medicine system on rational drug use in basic medical units in middle-western China." *Chinese Hospital Management*. 33(6): 41-43.
- Meng, Q. (2005). Review of Health Care Provider Payment Reforms in China: Background paper for World Bank China Rural Health Study (unpublished). Washington DC, World Bank.
- Ministry of Health Center for Health Statistics and Information (2009). *Research on Health Services of Primary Health Care Facilities in China, 2008 [in Chinese]*. Beijing, Peking Union Medical College Press.
- Peabody, J., R. Shimkhada, S. Quimbo, J. Florentino, M. Bacate, C. E. McCulloch and O. Solon (2011). "Financial incentives and measurement improved physicians' quality of care in the Philippines." *Health Aff (Millwood)* 30(4): 773-781.
- Savedoff, W. D., D. de Ferranti, A. L. Smith and V. Fan (2012). "Political and economic aspects of the transition to universal health coverage." *Lancet* 380(9845): 924-932.

- Sparrow, R., A. Suryahadi and W. Widyanti (2013). "Social health insurance for the poor: Targeting and impact of Indonesia's Askeskin programme." Social Science & Medicine 96(0): 264-271.
- Sun, Q., G. Zuo, K. Li, Q. Meng and P. He (2012). "Whether the essential medicine policy decrease the medical cost of rural population: experiences from three counties in Anhui Province." Chinese Health Economics 31(4): 65-67.
- Wagstaff, A., M. Lindelow, G. Jun, X. Ling and Q. Juncheng (2009a). "Extending health insurance to the rural population: an impact evaluation of China's new cooperative medical scheme." Journal of Health Economics 28(1): 1-19.
- Wagstaff, A., M. Lindelow, S. Wang and S. Zhang (2009b). Reforming China's Rural Health System. Washington DC, World Bank.
- Wagstaff, A., W. Yip, M. Lindelow and W. C. Hsiao (2009c). "China's health system and its reform: a review of recent studies." Health Economics 18 Suppl 2: S7-23.
- Wang, H., L. Zhang, W. Yip and W. Hsiao (2011). "An experiment in payment reform for doctors in rural China reduced some unnecessary care but did not lower total costs." Health Aff (Millwood) 30(12): 2427-2436.
- Witter, S., A. Fretheim, F. L. Kessy and A. K. Lindahl (2012). "Paying for performance to improve the delivery of health interventions in low- and middle-income countries." Cochrane Database Syst Rev 2: CD007899.
- World Health Organization (2010). World health report: health systems financing: the path to universal coverage. Geneva, WHO.
- Yang, H., Q. Sun, G. Zuo, K. Li and Q. Meng (2012). "Usage and structure of township hospitals under essential medicine system: cases in three counties of Shandong Province." Chinese Health Economics 31(4): 59-61.
- Yip, W. C., W. Hsiao, Q. Meng, W. Chen and X. Sun (2010). "Realignment of incentives for health-care providers in China." Lancet 375(9720): 1120-1130.
- Yip, W. C., W. C. Hsiao, W. Chen, S. Hu, J. Ma and A. Maynard (2012). "Early appraisal of China's huge and complex health-care reforms." Lancet 379(9818): 833-842.