BMJ Open

# Assessing dose–response effects of national essential medicine policy in China: comparison of two methods for handling data with a stepped wedge-like design and hierarchical structure

Yan Ren,[1] Min Yang,[1,2,3,4] Qian Li,[2] Jay Pan,[1,2] Fei Chen,[1] Xiaosong Li,[1] Qun Meng[5]

**To cite:** Ren Y, Yang M, Li Q, *et al*. Assessing dose–response effects of national essential medicine policy in China: comparison of two methods for handling data with a stepped wedge-like design and hierarchical structure. *BMJ Open* 2017;**7**: e013247. doi:10.1136/bmjopen-2016-013247

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2016-013247).

Received 30 June 2016
Revised 11 January 2017
Accepted 2 February 2017

For numbered affiliations see end of article.

**Correspondence to**
Dr Min Yang; yangmin2013@scu.edu.cn

## ABSTRACT

**Objectives:** To introduce multilevel repeated measures (RM) models and compare them with multilevel difference-in-differences (DID) models in assessing the linear relationship between the length of the policy intervention period and healthcare outcomes (dose–response effect) for data from a stepped-wedge design with a hierarchical structure.

**Design:** The implementation of national essential medicine policy (NEMP) in China was a stepped-wedge-like design of five time points with a hierarchical structure. Using one key healthcare outcome from the national NEMP surveillance data as an example, we illustrate how a series of multilevel DID models and one multilevel RM model can be fitted to answer some research questions on policy effects.

**Setting:** Routinely and annually collected national data on China from 2008 to 2012.

**Participants:** 34 506 primary healthcare facilities in 2675 counties of 31 provinces.

**Outcome measures:** Agreement and differences in estimates of dose–response effect and variation in such effect between the two methods on the logarithm-transformed total number of outpatient visits per facility per year (LG-OPV).

**Results:** The estimated dose–response effect was approximately 0.015 according to four multilevel DID models and precisely 0.012 from one multilevel RM model. Both types of model estimated an increase in LG-OPV by 2.55 times from 2009 to 2012, but 2–4.3 times larger SEs of those estimates were found by the multilevel DID models. Similar estimates of mean effects of covariates and random effects of the average LG-OPV among all levels in the example dataset were obtained by both types of model. Significant variances in the dose–response among provinces, counties and facilities were estimated, and the 'lowest' or 'highest' units by their dose–response effects were pinpointed only by the multilevel RM model.

**Conclusions:** For examining dose–response effect based on data from multiple time points with hierarchical structure and the stepped wedge-like designs, multilevel RM models are more efficient, convenient and informative than the multilevel DID models.

## Strengths and limitations of this study

- This study contributes new knowledge and experience in the choice of advanced multilevel difference-in-differences and multilevel repeated measures models to assess dose–response policy effects and variation of such effects among implementing units based on hierarchically structured panel data under a stepped wedge-like design.
- All model estimates are based on extensive national data for reliable and robust conclusions.
- The example data are not from a real stepped wedge design; potential differences in model estimates between a real stepped wedge and stepped wedge-like designs were not measured.

## INTRODUCTION

The most widely used method for assessing the effects of policy intervention in the fields of econometrics and sociology research with quantitative methods is the standard difference-in-differences (DID) analysis. Such analysis attempts to mimic an experimental research design using observational study data.[1] Based on data from two time points to test the difference in the differences over time between the treatment and control groups, the DID analysis measures excess changes due to exposure to treatment in comparison to change due to time effect alone in an untreated group. Application of the method can be found in assessing the impact of a policy on NHS dental check-ups in Scotland using data from the British Household Panel Survey,[2] in examining the

effects of the New Cooperative Medical Schemes (NCMS) on reducing the household's economic burden of chronic disease in rural China,[3] in examining the effects of the NCMS on child mortality, maternal mortality and school enrolment of 6–16 year- olds in China,[4] and in assessing the effects of policy change in the reimbursement rate on usage of outpatient, inpatient and pharmacy services in Veterans Affairs patients.[5] In all these applications, the conventional DID analysis was effective in measuring the overall effects of the intervention, regardless of the change of such effects over time, typically from two cross-sectional surveys or a panel study with data from two time points.

It is common for impacts of some policies to occur long after the intervention, so a longer period of follow-up with multiple time data points would be necessary to observe the effects that manifest gradually. Interrupted time series and segment regression analysis have been used to examine such changes of policy effects based on monthly data over several years.[6]

Conventional DID analysis assumes independence in data or no clustering effects. However, assessing policy effects often uses data from large-scale surveys with respondents clustered, such as individuals nested within a household, and households nested within regions,[7] or healthcare workers nested within healthcare facilities.[8] Such a hierarchical structure can violate the assumption of independence in data required by the conventional DID analysis because respondents from the same cluster tend to react in the same way to an intervention. Recently the DID analysis has been advanced to deal with data with a hierarchical structure. For example, Grytten *et al*[9] used multilevel DID models to examine the effects of a per capita-based remuneration system of dentists on the quality of dental care, and Arrieta[10] used the same model to assess the impact of Massachusetts healthcare reform on unpaid medical bills.

As current methods are effective in measuring the mean effects of a policy intervention during a certain period or the effects in changes over the period at the population level, none of them explore the variation in any impact that might be attributable to the contextual effects of socio-demographics, subculture environment and local policies. Identifying variation of effects beyond the main effects in the population could be very informative to policymaking in targeting local implementation. In recent years, the National Health and Family Planning Commission of China has established some surveillance information systems, such as the Health Statistical Information Center, the National Maternal and Child Health Routine Reporting System, and the National Maternal Mortality Surveillance System. Data from the national surveillance systems are extensive and available specifically for assessing policy implementation and evaluation of policy effects.

In the real world, particularly in developing countries, implementation of a national policy often starts with a small proportion of targeted individuals or facilities to pilot the first time block, then the policy is rolled out sequentially to involve more and more targets for the second, third and later time blocks until 100% of the targets are engaged in implementation.[11 12] Such inclusion of targets by time blocks is rarely by random allocation but according to convenience or purpose. Consequently, the data structure looks like a stepped-wedge design,[13] in which targets are grouped into time blocks reflecting the amount of exposure to the policy intervention. Such a case can be seen in establishing and implementing the National Essential Medicine Policy (NEMP) in China,[14 15] as well as the NCMS,[16] the separate two-child policy[17] and the 9-year compulsory education in China.[18]

The NEMP is an essential part of the national healthcare reform, aimed at improving availability, affordability, quality and safety of essential medicine. It requested government-run primary healthcare facilities (PHF), including urban-based community health centres, rural-based township and town centre hospitals to use low-cost medicine with zero profit.[19] Hence the same medicines described in PHFs were available at a lower cost than they were before by removing the income revenue from selling drugs in PHFs, and they were cheaper to get from PHFs than from general hospitals. A direct consequence was for PHFs to improve drug affordability, availability and rational use, as well as to attract more patients to their services. The NEMP was started in 2009, with no facilities implemented in 2008, and 27%, 26%, 25% and 22% of facilities were exposed to the policy each year from 2009 to 2012 respectively until all facilities implemented the policy. Thus facilities were exposed to the NEMP during different time periods and repeatedly reported data from 1 to 4 years by 2012. Such data were also presented in a hierarchical structure: facilities were nested within county and counties were nested within the province.

Based on such data, a previous study only examined the overall effects of the China NEMP in reducing costs of medicine and increasing service use by using the DID analysis with a PSM method.[20] The DID regression was often preceded by a PSM in an observational study. Accounting for the potential bias due to the non-random assignment of the policy, PSM is one strategy that corrects for selection bias in making estimates. The analysis ignored the hierarchical structure in the dataset and did not investigate the dose–response effect of the NEMP.

In this study we aimed to introduce advanced models that took into account both the hierarchical structure and the exposure time to a policy intervention in the dataset to answer the following three questions on policy effects that were not answered by the conventional DID method : (1) whether the expected overall effects changed as the implementation time increased, that is, a dose–response effect; (2) how much variation in the dose–response effect was attributable to context effects of provinces, counties and facilities respectively; and (3) how could we identify the best and worst performers in the policy implementation for effective management. We explored both multilevel DID and multilevel

repeated measures (RM) models to answer these questions. The concepts, specifications and interpretation of the two methods are illustrated and discussed.

## METHODS

### Example data and variables

From the national surveillance system on the NEMP in China, we extracted the total number of outpatient visits (OPVs) to facilities per year in 5 years (2008–2012) from 34 506 primary healthcare facilities (PHFs) of 2675 counties in 31 provinces in China to assess the policy effects on the service uses. The dataset forms a typical four-level hierarchical structure. A detailed description of the NEMP contents can be found elsewhere.[20] Data collection was administered by the Center for Information and Statistics and the National Health and Family Planning Commission of China. The PHF consisted of township hospitals (THs), central town hospitals, and community health centres according to their locality and resources. Context information on counties and provinces was obtained from the China Statistical Yearbook.[21]

The following variables were extracted for the purpose of illustration of the models.

- The dependent variable was the absolute number of OPVs to PHFs in total per year as a measure of service utilisation. It has been logarithmically transformed due to its skewed distribution of the raw scale and termed as LG-OPV.
- Time variables were two, one to mark the number of exposure years to policy implementation of a PHF (Exposure_t) and coded in the range 0–4, and one categorical variable of five levels (year) to indicate the data collection time in any year from 2008 to 2012.
- Potential covariates to the utilisation of the service were the ratio of health professionals to overall staff (RATIO_HP), the ratio of beds per staff (RATIO_B) (grouped as <5, 5–9, 10–20, >20), log(ratio of total assets per staff) (LG_RTA), the type of facility (ToF) to capture community health service centres (CHCs), town central hospitals (TCHs) and THs. At the provincial level, one variable (region) indicated eastern, central and western China.

The means and SDs of the raw dependent variable by level of covariates are presented in table 1. We can observe the highest OPVs in urban-based facilities (CHCs) and those in eastern China compared with their counterparts. The means for each raw dependent variable at any variable level demonstrate an increased number of OPVs from 2008 to 2012 to reflect the effect of year. The means on the diagonal between the policy exposure years and the calendar year of policy implementation show a linear increase to suggest some degree of dose–response effect ; that is, the more time taken to implement the NEMP by facilities, the more OPVs to the facilities. The model analysis attempts to establish statistical evidence for these phenomena, and further to find contextual variation in the changing patterns

**Table 1** General description of the absolute number of outpatient visits (OPVs) to facilities stratified by facility type, exposure time, regions and year

| | Years of NEMP implementation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2008 | | 2009 | | 2010 | | 2011 | | 2012 | |
| Variable | Facilities | Mean (SD) | Facilities | Mean (SD) | Facilities | Mean (SD) | Facilities | Mean (SD) | Facilities | Mean (SD) |
| Facility type | | | | | | | | | | |
| CHC | 2218 | 62 766 (89 557) | 2218 | 73 065 (104 648) | 2218 | 80 821 (116 769) | 2218 | 86 024 (120 366) | 2218 | 93 200 (125 587) |
| TH | 22 777 | 21 713 (31 555) | 22 777 | 23 619 (36 573) | 22 777 | 24 249 (38 634) | 22 777 | 25 139 (41 127) | 22 777 | 28 342 (43 945) |
| TCH | 9511 | 21 674 (31 434) | 9511 | 23 521 (35 047) | 9511 | 24 035 (35 982) | 9511 | 25 013 (37 859) | 9511 | 27 795 (41 178) |
| Policy exposure years | | | | | | | | | | |
| 0 | 34 506 | 24 341 (39 323) | 25 240 | 27 192 (49 161) | 16 305 | 27 861 (53 988) | 7715 | 27 456 (58 055) | 7715 | 30 540 (60 985) |
| 1 | | | 9266 | 25 620 (33 647) | 8935 | 29 186 (51 971) | 8590 | 31 349 (57 955) | 8590 | 34 971 (62 014) |
| 2 | | | | | 9266 | 26 456 (35 243) | 8935 | 29 966 (52 149) | 8935 | 33 143 (54 243) |
| 3 | | | | | | | 9266 | 27 244 (37 070) | 9266 | 30 700 (41 202) |
| 4 | | | | | | | | | | |
| Regions | | | | | | | | | | |
| Eastern | 8983 | 41 188 (60 800) | 8983 | 46 154 (70 298) | 8983 | 49 213 (76 038) | 8983 | 52 387 (79 825) | 8983 | 57 495 (84 249) |
| Central | 10 375 | 21 783 (30 998) | 10 375 | 23 048 (34 012) | 10 375 | 23 562 (34 402) | 10 375 | 24 110 (34 173) | 10 375 | 27 544 (36 474) |
| Western | 15 148 | 16 103 (20 780) | 15 148 | 17 825 (26 185) | 15 148 | 18 065 (29 569) | 15 148 | 18 522 (32 302) | 15 148 | 20 754 (34 797) |
| Total | 34 506 | 24 341 (39 323) | 34 506 | 26 770 (45 523) | 34 506 | 27 827 (49 103) | 34 506 | 29 018 (51 630) | 34 506 | 32 360 (54 861) |

The eastern region includes 11 provinces, the central region includes 8 and the western includes 12 provinces according to the National Statistics Year Book of China (2015).
CHC, community health centre; NEMP, national essential medicine policy; TCH, town centre hospital; TH, township hospital.

beyond the main effects of the policy. The description of the main independent variables is presented in online supplementary appendix table S1.

### Study design

In 2008 no PHF had implemented the NEMP intervention. In 2009 about 27% of PHFs had implemented the policy, which was scaled up gradually to cover about 53% of PHFs in 2010, 78% in 2011 and 100% in 2012. This means that the 9266 facilities started in 2009 had 4 years' exposure to the policy intervention by the end of 2012; 8935 facilities started in 2010 had 3 years' exposure; 8590 facilities started in 2011 had 2 years' exposure; and 7715 facilities started in 2012 had 1 year's exposure.

In the context of a cluster randomised trial, the stepped-wedge design involves the collection of observations during a baseline period in which no clusters are exposed to the intervention. Following this, at regular intervals, or steps, a cluster (or group of clusters) is randomised to receive the intervention. This process continues until all clusters have crossed over to receive the intervention. Observations are taken at every cluster and at each period. Stepped-wedge studies typically have one period in which observations are made while all clusters are unexposed to the intervention, and one period in which all clusters are exposed to the intervention. A number of observations in each cluster and period made up the data structure.[13] Because in reality the PHFs assigned to implement the NEMP policy each year were not random but selected by administrative convenience, the data structure presented in table 2 shows a stepped wedge-like design with four steps.

## MODELS

### Conventional DID and multilevel DID models

The basic conventional DID model for the overall effect of a policy intervention is written as M1:

$$y_i = \beta_0 + \beta_1(\text{time})_i + \beta_2(\text{intervention})_i + \beta_3(\text{time} \times \text{intervention})_i + e_i \quad \text{(M1)}$$

$$e_i \sim N(0, \sigma^2)$$

In the model, i (=1, 2, …, n) indicates facilities in our case. The parameter $\beta_3$ estimates the mean effect of the policy, which is the difference in the differences between the treatment and control groups over two time points, before and after the intervention, hence the difference in differences. To add other covariates such as region and facility type for subgroup effects in the model is straightforward.[22] Clearly, a significant strength of this method is that time effects, and unobserved time-invariant confounders are removed from the estimation. However, the accuracy of the DID method depends on one critical assumption: time effects have to be the same for both the intervention and control groups, so the composition of the intervention and the control group must be the same over time. For non-random observational data, significant imbalances in characteristics between control and intervention PHFs exist before policy implementation; simple multivariate regression may not be powerful enough to adjust for the imbalances between comparison groups. PSM between the control and intervention groups on some key covariates before fitting the DID models has been a widely used approach to deal with the imbalance issue.[23] The most commonly used matching method is nearest neighbourhood matching.[24] In this method, the cases and control units are randomly sorted, and then the cases sequentially matched to the nearest unmatched control even if the absolute difference values of the propensity score between the selected case and the control under consideration are not close. To acquire good matches, the greedy matching techniques were used to create a propensity score matched pair sample using a user-written SAS macro,[25] which demands that the absolute difference in the propensity score of the case and the control is small, begins with a smaller difference (such as 0.00001) to match, and then gradually the difference is increased to 0.1. Using this method, we first fitted a logistic regression to estimate the propensity scores based on the variables ratio of health professionals to overall staff, ratio of beds per staff, LG_RTA, type of facility (CHC, TCH and TH), and region (eastern, central and western China). Then the cases were ordered and sequentially matched to the nearest unmatched control within a certain range of difference. If more than one unmatched control was matched to a case, the control was selected at random. Once a match was made, the match was not reconsidered. Only matched control and treatment units were included in the DID modelling analysis.

Given 5 years of panel data we wanted to assess the possible linear change in the dependent variable according to the length of policy implementation in years; a dose–response relationship can be captured by defining a continuous variable for the length of policy exposure years so that the dependent variable is a function of the policy exposure variable. However, the time variable in M1 is a binary term indicating a difference in only two time points. We needed to construct four models based on the common control and intervention group but different exposure times, that is, 1 year, 2 years and so on to estimate policy intervention effects corresponding to the different lengths of policy exposure period. Table 3 shows the structure of the example data for estimating the dose–response effects by the DID models. We defined 2008 as before intervention time 1, and 2009, 2010, 2011 and 2012 as after intervention time 2. Facilities that implemented the NEMP in 2009 formed the intervention group and were intervened for 4 years by 2012. Similarly, facilities that implemented the NEMP in 2010 formed the intervention group and were intervened for 3 years by 2012. Those that implemented the NEMP in 2011 were intervened for 2 years by 2012. Facilities that implemented the NEMP in 2012 formed the control group and were

Table 2 Stepped wedge data structure

| | Time periods | | | | |
|---|---|---|---|---|---|
| | 2008 | 2009 | 2010 | 2011 | 2012 |
| PHF clusters (exposure) | | | | | |
| 4 | **9266 (27%)** | *9266 (27%)* | *9266 (27%)* | *9266 (27%)* | *9266 (27%)* |
| 3 | **8935 (26%)** | **8935 (26%)** | *8935 (26%)* | *8935 (26%)* | *8935 (26%)* |
| 2 | **8590 (25%)** | **8590 (25%)** | **8590 (25%)** | *8590 (25%)* | *8590 (25%)* |
| 1 | **7715 (22%)** | **7715 (22%)** | **7715 (22%)** | **7715 (22%)** | *7715 (22%)* |

Italic text represents intervention periods; bold text represents control periods. Each entry represents a data collection point. Numbers 1–4 represent the length of exposure time to the NEMP intervention in years.
NEMP, national essential medicine policy; PHF, primary healthcare facility.

intervened for only 1 year. Without PSM, the sample sizes for intervention and control were 9266 and 7715, respectively. Using nearest-neighbourhood matching with the absolute difference value of the propensity score between the case and the control group from 0.00001 to 0.1, the final sample sizes for intervention and control facilities were 7393 and 7393, respectively, as shown in table 3.

One critical limitation arose from using M1 in analysing the example data that were clustered at several levels, and each level could have common context factors shared among the units. For example, facilities in the same county could have more similar outcomes than those from other counties, which brings dependence in the data due to clustering. Consequently, statistical estimates of this model could be biased.[26 27]

To overcome this critical limitation, either correcting SEs of regression coefficients using robust statistics or multilevel DID models could be considered. We chose multilevel DID models for a three-level structure in the example data: facilities at level 1, county at level 2 and province at level 3. Following the definition of the MLwiN software,[28] the letters i, j, k indicate units at levels 1, 2 and 3, respectively, and a basic three-level DID model could be written as follows:

$$y_{ijk}=\beta_{0ijk}+\beta_1(\text{time})_{ijk}+\beta_2(\text{intervention})_{ijk}+\beta_3(\text{time}\times\text{intervention})_{ijk} \quad \text{(M2)}$$

$$\beta_{0ijk}=\beta_0+v_{0k}+u_{0jk}+e_{0ijk}$$

$$v_{0k}\sim N(0,\sigma^2_{v0}),\ u_{0k}\sim N(0,\sigma^2_{u0}),\ e_{0ijk}\sim N(0,\sigma^2_{e0})$$

$$\text{cov}(v_{0k},u_{0jk},e_{0ijk})=0$$

$$\text{var}(y_{ijk})=\sigma^2_{v0}+\sigma^2_{u0}+\sigma^2_{e0}$$

This model has four partial regression coefficients, the same as in the conventional DID model, M1. The partial regression coefficients under multilevel model analysis are termed as *fixed* partly because they estimate the mean effects of covariates that are the same or fixed for every individual or facility, in this case in the model. In M2 the overall mean policy effect between any two time points is estimated and tested by the regression coefficient $\beta_3$, which is the same for all facilities. In contrast to the *fixed* part in M2, the *random part* in the model is composed of parameters$\sigma^2_{v0}$, $\sigma^2_{u0}$ and $\sigma^2_{e0}$ for variation of the intercept or the overall mean estimate of the dependent variable among provinces, counties and facilities. In this case the random variable $v_{0k}$represents the difference between the mean of the kth province and the overall mean $\beta_0$, and the random variable $u_{0jk}$ represents the difference between the jth county mean within the kth province and the grand $\beta_0$. They are also termed random effects and assumed to come from normal distribution. Random effects could be attributed to context effects such as socioeconomics or health systems or healthcare resources or local policy for the healthcare of the provinces or counties.

Clearly the model M2 effectively deals with clustering effects in the data by including random effects in the outcome at each level of the hierarchy; the context effects of the outcome can be examined by levels. Other covariates can be added to M2 straightforwardly.

To assess the dose–response effects of the NEMP using M2, we also need to construct four models as mentioned previously for the model M1. The modelling process for a simple linear relationship under this context becomes cumbersome and ineffective. Consequently for a set of point estimates from different models for the dose–response effects of the policy, there was not an easy approach to quantify whether such policy effects varied among provinces or counties or facilities.

## Multilevel RM models

Based on the study design which presents cumulative exposure to the policy and the clustering feature, we propose the following four-level RM regression models that take into account the clustering effects in the data while assessing the dose–response effect of the policy and estimating the random effects of the dose–response effect across provinces, counties and

**Table 3** Data construction for potential dose–response effect analysis using multilevel DID model (M1 and M2)

| | 2008 time 1 Before intervention | | 2009 time 2 After intervention | | 2010 time 2 After intervention | | 2011 time 2 After intervention | | 2012 time 2 After intervention | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group | Facilities | Exposure time (year) | Facilities | Exposure time (year) | Facilities | Exposure time (year) | Facilities | Exposure time (year) | Facilities | Exposure time (year) |
| Intervention | 7393 | 0 | 7393 | 1 | 7393 | 2 | 7393 | 3 | 7393 | 4 |
| Control | 7393 | 0 | 7393 | 0 | 7393 | 0 | 7393 | 0 | 7393 | 1 |

The intervention group represents the facilities that implemented the NEMP in 2009 and intervened for 4 years by 2012.
The control group represents the facilities that implemented the NEMP in 2012 and intervened for 1 year by 2012.
DID, difference-in-differences; NEMP, national essential medicine policy.

facilities. A facility in the four-level structure becomes a level 2 unit above repeated measure in time which is the level 1 unit. This involves a change in the level indicators in the above models. To answer our research questions, three consecutive four-level models are constructed.

Following the definition of the software MLwiN for multilevel models, the letters i, j, k and l denote the repeated measure in time (level 1), facility (level 2), county (level 3) and province (level 4), respectively. The following model estimates an overall linear effect with exposure time of policy measured by parameter $\beta_1$ and partitions the total variance in the outcome into four components for the four levels of sources: $\sigma^2_{w0}$, $\sigma^2_{v0}$, $\sigma^2_{u0}$ and $\sigma^2_{e0}$ for the random effects among provinces, counties, facilities and repeated measure in time, respectively. The assumptions of independence and uncorrelated relationship over the random effects at different levels are the same as for M2:

$$y_{ijkl}=\beta_{0ijkl}+\beta_1(\text{exposure\_t})_{ijkl} \quad \text{(M3)}$$

$$\beta_{0ijkl}=\beta_0+w_{0l}+v_{0kl}+u_{0jkl}+e_{0ijkl}$$

$$w_{0l}\sim N(0,\sigma^2_{w0}), v_{0kl}\sim N(0,\sigma^2_{v0}), u_{0jkl}\sim N(0,\sigma^2_{u0}), e_{0ijkl}\sim N(0,\sigma^2_{e0})$$

$$\text{cov}(w_{0l,}v_{0kl},u_{0jkl},e_{0ijkl})=0$$

$$\text{var}(y_{ijkl})=\sigma^2_{w0}+\sigma^2_{v0}+u^2_{u0}+e^2_0$$

In this model the parameter $\beta_1$ is a unit of linear change in the dependent variable for every year of the intervention period of the policy, the same interpretation of slope as in any regression model. To separate the dose–response effects of the policy from the effects due to the time change in a calendar year in healthcare conditions and human resources of the facilities we added some covariates to the fixed part of M3 with everything else in the model unchanged:

$$y_{ijkl}=\beta_{0ijkl}+\beta_1(\text{exposure\_t})_{ijkl}+\sum_{h=1}^{4}\beta_{2h}(\text{year})_{hijkl}+\sum\beta_{3f}(X_f)_{jkl} \quad \text{(M4)}$$

$$\beta_{0ijkl}=\beta_0+w_{0l}+v_{0kl}+u_{0jkl}+e_{0ijkl}$$

The X denotes a set of covariates, such as beds per staff and assets of facilities, or facility type, regions and so on. The time differences in years, that is, 2009 vs 2008, 2010 vs 2008, are estimated by parameters $\beta_{2h}$ and the dose–response effect of policy is estimated by the parameter $\beta_1$ The latter is independent of the calendar time effects.

Further, to find evidence on whether the policy effects were different by province, by county and by facility, we

assumed random effects of the parameter $\beta_1$ in the following models:

$$y_{ijkl}=\beta_{0ijkl}+\beta_{1jkl}(\text{exposure\_t})_{ijkl}+\sum_{h=1}^{4}\beta_{2h}(\text{year})_{hijkl}+\sum\beta_{3f}(X_f)_{kjl} \quad (M5)$$

$$\beta_{0ijkl}=\beta_0+w_{0l}+v_{0kl}+u_{0jkl}+e_{0ijkl}$$

$$\beta_{1jkl}=\beta_1+w_{1l}+v_{1kl}+u_{1jkl}$$

$$(w_{0k},w_{1k}) \sim MN(0,\Omega_w),(v_{0kl},v_{1kl}) \sim MN(0,\Omega_v),(u_{0jkl},u_{1jkl}) \sim MN(0,\Omega_u)$$

$$(e_{0ijkl}) \sim N(0,\sigma^2_{e0})$$

$$\Omega_w=\begin{pmatrix}\sigma^2_{w0} & \\ \sigma_{w01} & \sigma^2_{w1}\end{pmatrix},\Omega_v=\begin{pmatrix}\sigma^2_{v0} & \\ \sigma_{v01} & \sigma^2_{v1}\end{pmatrix},\Omega_u=\begin{pmatrix}\sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1}\end{pmatrix}$$

Four sets of random effects of policy implementation are presented in M5. The model has two random effects at each of the province, county and facility levels. The terms $w_{0l},v_{0kl},u_{0jkl}$ are random effects of the overall mean, and the terms $w_{1l},v_{1kl},u_{1jkl}$ are the dose–response random effects at each of the above levels accordingly. By estimating and testing variance terms $\sigma^2_{w1}$, $\sigma^2_{v1}$and $\sigma^2_{u1}$ of the random effects $w_{1l},v_{1kl},u_{1jkl}$, respectively, we can find out the distribution of the dose–response random effects at a different level and identify the 'best' or 'worst' units (province or county or facility) in the policy implementation for further management. The latter task can be achieved by estimating and ranking random effects $w_{1l},v_{1kl}$ and $u_{1jkl}$.

While the multilevel DID model (M2) used the data presented in table 3, the multilevel RM models (M3–M5) used the full data of all facilities from 2008 to 2012, as shown in table 2.

### Model fitting and comparison

Multilevel models for data under a three-level hierarchy can be fitted by the usual software such as SAS or Stata. For our data with a four-level hierarchy, we used MLwiN 2.30 for all modelling analysis and SAS 9.3 for descriptive analysis.

We used the Wald statistic to test the significance of fixed effects estimated by $\beta_s$ and random effects estimated by variances at different levels. To compare the quality of fit between nested models, we used the −2LogLikelihood (−2LL) value. The smaller the −2LL value, the better the model.

## RESULTS

To examine the cumulative effects of NEMP exposure over time, four three-level DID models defined by M2 were fitted with the same controls to reflect exposure in years 1, 2, 3 and 4, respectively, and matched using propensity scores. We expected to observe an overall policy effect from each of the four models and also possible dose–response effects of the NEMP based on these models. Meanwhile, the four-level RM models defined by M4 and M5 were fitted for the dose–response effect and variation of such effects among provinces, counties and facilities. The results of the two methods were compared using the following aspects.

### Dose–response effect

From a series of four multilevel DID models adjusted for the covariates, a significant policy effect of increased LG-OPV was estimated as 0.036 at 1 year of policy exposure, 0.073 at 2 years' exposure, 0.068 at 3 years' and 0.097 at 4 years' exposure. Assuming a linear increase in the policy effects and a total cumulative effect at 0.061 (from 0.036 to 0.097) over the 4-year period, a dose–response quantity would be roughly estimated as 0.015 per year of exposure to the policy. However, when we fit the multilevel RM model (M4), a significant dose–response effect was estimated directly at 0.012 for every year of exposure to the policy and 0.048 of total cumulative effects over the 4-year period (table 4). Although the linear effects over the policy exposure periods estimated by the two models are close, a much smaller SE of the effect estimate from the multilevel RM model was observed than the former models (0.002 vs 0.018) due to the fact that the multilevel DID models did not include all facilities in the calculation; namely the data size was smaller than that for the multilevel RM model. Consequently, the U value of the dose–response effect estimate from the multilevel RM model was bigger than that for the four multilevel DID models. This suggests that the estimated policy effects from the latter model have higher statistical efficiency than those from the former models. Also, one estimate of a dose–response effect from one model is certainly much more straightforward and efficient than from four models with indirect estimation.

### Random effects

Two random effects in our case are present: random effects related to the overall mean of the LG-OPV, which are also termed random intercepts, and random effects related to the dose–response effects of the policy effect, which are also termed random slopes. Random effects are summarised by estimates of the corresponding variance in our models. In table 5, both multilevel DID and multilevel RM models produced significant variances on the overall mean of LG-OPV across the province and county, and facility. Such results suggest that multilevel modelling is necessary in appropriate data with hierarchical structure, and in this case the multilevel DID

**Table 4** Estimates of the dose–response effects of the NEMP on LG-OPV

| Method | coefficient | | Intervention year (exposure time to NEMP implementation in years) 2009 (1) | 2010 (2) | 2011 (3) | 2012 (4) |
|---|---|---|---|---|---|---|
| Multilevel DID (M2) | Intercepts | $\beta_0$ (SE) | 7.859 (0.170) | 7.889 (0.171) | 8.039 (0.171) | 8.208 (0.173) |
| | | U value (p) | 46.23 (<0.0001) | 46.13 (<0.0001) | 47.01 (<0.0001) | 47.45 (<0.0001) |
| | Time × intervention (effect of the NEMP) | $\beta_3$ (SE) | 0.036 (0.018) | 0.073 (0.018) | 0.064 (0.018) | 0.097 (0.019) |
| | | U value (p) | 2.00 (0.044) | 4.06 (<0.0001) | 3.56 (0.0005) | 5.11 (<0.0001) |
| Multilevel RM (M4) | Intercepts | $\beta_0$ (SE) | 9.610 (0.179) | | | |
| | | U value (p) | 53.69 (<0.0001) | | | |
| | Exposure times | $\beta_1$ (SE) | 0.012 (0.002) | | | |
| | | U value (p) | 6.00 (<0.0001) | | | |

DID, difference-in-differences; NEMP, national essential medicine policy; RM, repeated measures.

**Table 5** Estimates of the random effects of the NEMP on LG-OPV

| Method | Random effects | Intervention year | Province $\sigma^2$ (SE) | Province $ICC_{province}$ | Province 95%CI (p) | County $\sigma^2$ (SE) | County $ICC_{county}$ | County 95%CI (p) | Facility $\sigma^2$ (SE) | Facility 95%CI (p) | Time $\sigma^2$ (SE) | Time 95%CI (p) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Multilevel DID (M2) | Intercepts | 2009 | 0.305 (0.072) | 47.14 | 0.446 to 0.164 (<0.0001) | 0.342 (0.012) | 36.62 | 0.318 to 0.366 (<0.0001) | 0.592 (0.005) | 0.582 to 0.602 (<0.0001) | | |
| | | 2010 | 0.308 (0.073) | 47.83 | 0.165 to 0.451 (<0.0001) | 0.336 (0.012) | 35.52 | 0.312 to 0.360 (<0.0001) | 0.610 (0.005) | 0.600 to 0.620 (<0.0001) | | |
| | | 2011 | 0.314 (0.074) | 49.37 | 0.169 to 0.459 (<0.0001) | 0.322 (0.011) | 33.79 | 0.300 to 0.344 (<0.0001) | 0.631 (0.005) | 0.621 to 0.641 (<0.0001) | | |
| | | 2012 | 0.324 (0.076) | 50.23 | 0.175 to 0.473 (<0.0001) | 0.321 (0.011) | 33.30 | 0.299 to 0.343 (<0.0001) | 0.643 (0.006) | 0.631 to 0.655 (<0.0001) | | |
| Multilevel RM (M5) | Intercepts | 2009–2012 | 0.397 (0.103) | 55.92 | 0.195 to 0.599 (0.0001) | 0.313 (0.011) | 35.09 | 0.291 to 0.335 (<0.0001) | 0.579 (0.005) | 0.569 to 0.589 (<0.0001) | 0.152 (0.001) | 0.150 to 0.153 (<0.0001) |
| | NEMP effect | | 0.003 (0.001) | 21.43 | 0.001 to 0.005 (0.0003) | 0.011 (0.000) | 33.33 | 0.010 to 0.012 (<0.0001) | 0.022 (0.000) | 0.023 to 0.022 (<0.0001) | | |

DID, difference-in-differences; NEMP, national essential medicine policy; RM, repeated measures.

models are more appropriate than the conventional DID models. The sizes of variances estimated at each of the three data structure levels by both models are similar; for example, the percentage of the variations from the multiple RM models among provinces, counties and facilities is 30.8%, 24.3% and 44.9%, respectively, while that of the multilevel DID 2012 model is 25.2%, 24.9%, and 50%, respectively. The intra-province and intra-county correlation coefficients from the multiple RM models are 55.9% and 35.1%, respectively, while those from the multilevel DID 2012 model is 50.2% and 33.3%, respectively.

The small differences in the estimates are due to different sample sizes used in those models, and also to the fact that the multilevel DID models balanced out the year difference, and the multilevel RM models treat the year as RM and estimate a variance for the year differences as shown in table 5. Overall, the two models are comparable in estimating variances of random intercepts for provinces, counties and facilities.

However, the multilevel DID models cannot estimate variance related to the random effects of the NEMP effect. In table 5, the variance of the dose–response effects or random slopes of the NEMP effect at the province, county and facility levels were estimated respectively and simultaneously by the multilevel RM model M5. All estimated variances were statistically significant, which implies that the NEMP effect over the increased LG-OPV was variable among provinces, counties or facilities in China. Some provinces could have faster increases than others, and some might not increase at all. The same interpretation can be applied to counties and facilities. It is notable from the results of M5 that out of the total random slope variation, 8.3% was attributable to the difference among provinces, 30.6% to the difference among counties and 61.1% to facility differences. The intra-province and intra-county correlation coefficients are 21.4% and 33.3%, respectively. At this stage, identifying the best or worst performers of the policy implementation at any of the three levels might be necessary for examining which context variables could explain the variation in the policy effects at the corresponding levels.

### Identifying units with the best and worst policy effects

From the multilevel RM model M5, we can calculate the random slopes of the dose–response effects $w_{1l}$,$v_{1kl}$,$u_{1jkl}$ of the province, county and facility, respectively. For example, figure 1 shows the dose–response effects by province. It can be seen that the trend in the policy effect varied between provinces, with some having increased more, some less and some decreased. When we calculated the size of estimated random slopes of provinces, the dose–response effects of provinces $\beta_1+w_{1l}$ were marked on the map in figure 2. We can easily visualise the province with the best dose–response effect, that is, fast increasing LG-OVP with increased time to implement the NEMP, which is province A in eastern China, and with decreased LG-OVP, that is, the worst performer, province B in central China. The significant dose–response effect of province A was estimated directly as 0.14 increase for every 1 year of exposure to the policy, but province B showed a 0.097 decrease.

### Effect of time in calendar year

Each of the four multilevel DID models M2 included the parameter $\beta_1$ for the time effects of calendar years 2009, 2010, 2011 and 2012 in comparison to 2008. The results in table 6 show significantly increased LG-OPV from 2008 to 2012, with 2.55 times more OPV in 2012 than in 2009. The multilevel RM model treated calendar year as a set of covariates as defined in M4. The comparable results of multilevel RM models in table 6 also demonstrate a trend of increased LG-OPV from 2008 to 2012, with 2.51 times more OPV in 2012 than in 2009. Although both models gave a similar change trend due to the effects of the calendar year on the relative measure, the test values U of multilevel RM models are

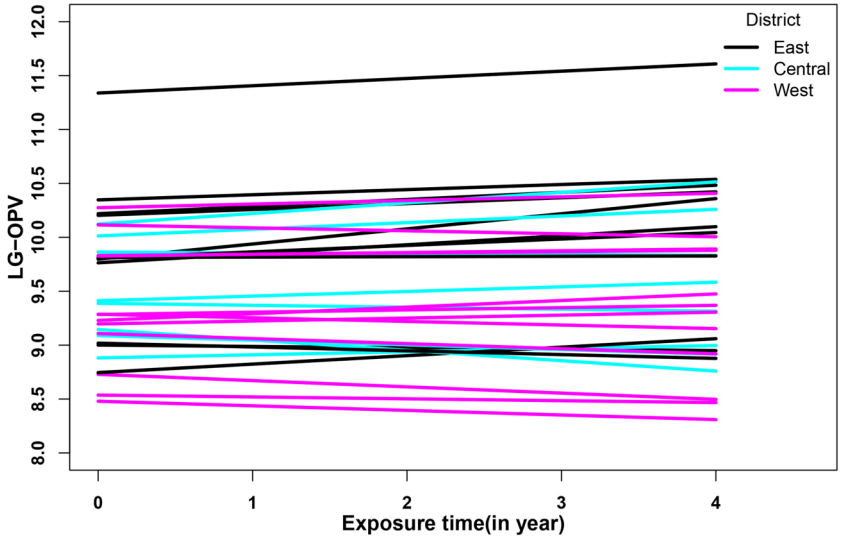


**Figure 1** Estimated dose–response effects on the logarithm number of outpatient visits per facility per year (LG-OPV) by provinces by region.

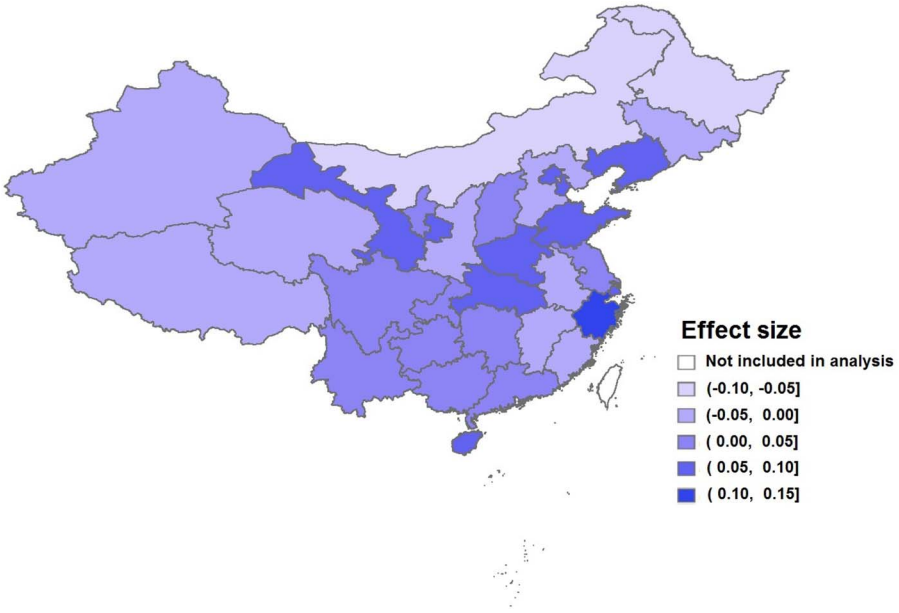


**Figure 2** The estimated dose–response effects by provinces.

consistently higher than those of multilevel DID models, therefore a more robust and efficient estimate of the time effects was shown by the former. By comparing parameter estimates among multilevel RM models with and without the year covariates, we found a decreased dose–response random coefficient at facility level from 0.064 to 0.014, which suggests a 78% variation in the dose–response effects at the facility level was explained by the year effects. We could not perform the same analysis using multilevel DID models.

### Effects of other covariates

The effects of facility level covariates on the LG-OVP were assessed by the four multilevel DID models and the multilevel RM model. The results shown in table 7 are consistent between the two types of models. The higher RATIO_HP and the transformed LG_RTA were significantly associated with more OPV, while the ratio of beds per staff (RATIO_B) was not related to the outcome. Compared with community health centres in urban settings, the rural-based township or town centre facilities had significantly lower OPV. Compared with facilities in eastern China, those in central and western China had significantly lower OPV, with facilities in western China at the lowest level. By comparing parameter estimates between multilevel RM models with and without these covariates, we found a minor decrease in the dose–response random coefficient at facility level from 0.014 to 0.012, which suggests that a 14% variation in the dose–response effects at the facility level was explained by those facility level variables.

## DISCUSSION

The implementation of the NEMP was part of the deepened healthcare reform in China since 2009. The main aims of the policy were to alleviate the burden on citizens of expensive medical bills and increase their access to healthcare services. A number of studies have evaluated the effects of the NEMP in PHFs after implementation of the NEMP, such as the change in drug prices before and after implementation,[29–31] availability and affordability of essential medicines,[19 29 31] rational use of essential medicines,[30 32–35] medicine expenditure,[20 33 36] outpatient service use[20] and so on. As a common practice, all the previous studies used the conventional DID in their evaluation, and all ignored clustering effects among counties or facilities over time. No study tried to explore the variability of the policy effects among facilities that could be explained by context factors. No study compared the multilevel DID model and multilevel RM model to evaluate the dose–response effect of the NEMP over a longer period of implementation. In this study we illustrated that when analysing data with a hierarchical structure, the multilevel DID model is more appropriate than the conventional DID model. We also showed that when data were collected at multiple time points with hierarchical structure under a stepped wedge-like design, the multilevel RM model is more appropriate, efficient and powerful than the multilevel DID models in assessing the dose–response effects of policy and variation of the effects among provinces or counties or facilities. We demonstrated and discussed the similarities and differences of the two models using one example.

For data with a hierarchical structure, as in our example, models with random effects such as multilevel models or mixed models have been widely accepted as appropriate tools for data analysis.[37] Conventional DID analysis ignores the dependence on the outcome measure due to clustering effects of the data structure and can seriously bias estimates of the intervention effects.[38] The significant random effects on the mean LG-OVP among provinces, counties and facilities

**Table 6** Estimates for time effects in calendar year on the LG-OPV

| Method | 2008 | 2009 | | 2010 | | 2011 | | 2012 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Est (SE) | U value (p) | Est (SE) | U value (p) | Est (SE) | U value (p) | Est (SE) | U value (p) |
| Multilevel DID (M2) | Control | 0.069 (0.013) | 5.31 (<0.0001) | 0.058 (0.013) | 4.46 (<0.0001) | 0.078 (0.013) | 6.00 (<0.0001) | 0.176 (0.014) | 9.00 (<0.0001) |
| Multilevel RM (M4) | R=Reference | 0.080 (0.003) | 26.67 (<0.0001) | 0.087 (0.004) | 21.75 (<0.0001) | 0.100 (0.005) | 20.00 (<0.0001) | 0.201 (0.007) | 28.71 (<0.0001) |

Est stands for parameter estimate by regression coefficients in each of the models. The estimates of time effects are contrasted to the year 2008.
DID, difference-in-differences; RM, repeated measures.

**Table 7** Estimates of effects of other covariates in association with LG-OPV

| Variables | Multilevel RM (M4) 2009 | | 2010 | | 2011 | | 2012 | | Multilevel RM (M4) 2008–2012 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Est (SE) | U value (p) | Est (SE) | U value (p) | Est (SE) | U value (p) | Est (SE) | U value (p) | Est (SE) | U value (p) |
| RATIO_HP | 1.288 (0.038) | 33.89 (<0.0001) | 1.282 (0.039) | 32.87 (<0.0001) | 1.283 (0.039) | 32.90 (<0.0001) | 1.286 (0.040) | 32.15 (<0.0001) | 0.312 (0.013) | 24.00 (<0.0001) |
| LG_RTA | 0.128 (0.005) | 25.60 (<0.0001) | 0.130 (0.005) | 26.00 (<0.0001) | 0.118 (0.004) | 29.50 (<0.0001) | 0.102 (0.004) | 25.50 (<0.0001) | 0.032 (0.001) | 32.00 (<0.0001) |
| RATIO_B (<5) (reference) | | | | | | | | | | |
| 5–9 | −0.068 (0.104) | −0.65 (0.511) | −0.047 (0.110) | −0.43 (0.670) | −0.067 (0.126) | −0.53 (0.594) | −0.125 (0.121) | −1.03 (0.303) | 0.027 (0.033) | −0.82 (0.414) |
| 10–20 | 0.027 (0.409) | 0.07 (0.950) | 0.045 (0.368) | 0.12 (0.904) | −0.177 (0.372) | −0.48 (0.635) | −0.019 (0.342) | −0.05 (0.956) | −0.046 (0.115) | −0.40 (0.691) |
| >20 | −2.920 (0.466) | −6.27 (<0.0001) | −0.958 (0.570) | −1.68 (0.093) | −0.644 (0.479) | −1.34 (0.178) | 0.079 (0.582) | −0.14 (0.890) | −0.133 (0.135) | −0.99 (0.327) |
| Facility type (CHC) (reference) | | | | | | | | | | |
| TH | −0.090 (0.030) | −3.00 (0.003) | −0.099 (0.030) | −3.30 (0.001) | −0.116 (0.031) | −3.74 (0.0001) | −0.113 (0.031) | −3.65 (0.0003) | −0.170 (0.025) | −6.80 (<0.0001) |
| TCH | −0.113 (0.031) | −3.65 (0.0003) | −0.121 (0.031) | −3.90 (0.0001) | −0.126 (0.031) | −4.06 (<0.0001) | −0.131 (0.032) | −4.09 (<0.0001) | −0.174 (0.026) | −6.69 (<0.0001) |
| Region (East) (reference) | | | | | | | | | | |
| Central | −0.603 (0.215) | −2.80 (0.005) | −0.621 (0.216) | −2.88 (0.032) | −0.635 (0.218) | −2.91 (0.004) | −0.603 (0.221) | −2.73 (0.006) | −0.631 (0.287) | −2.20 (0.028) |
| West | −1.024 (0.200) | −5.12 (<0.0001) | −1.064 (0.201) | −5.29 (<0.0001) | −1.080 (0.202) | −5.35 (<0.0001) | −1.079 (0.204) | −5.29 (<0.0001) | −0.809 (0.207) | −3.91 (<0.0001) |

Est stands for parameter estimate of each covariate by regression coefficients in each of the models.
CHC, community health centre; LG_RTA, log(ratio of total assets per staff); RATIO_B, ratio of beds per staff (grouped as <5, 5–9, 10–20, >20); RATIO_HP, ratio of health professional over all staff; RM, repeated measures; TCH, town centre hospital; TH, township hospital.

estimated by both multilevel DID and RM models in our analysis suggested strong clustering effects in the example data, which cannot be dealt with by conventional DID analysis. Hence multilevel models as an advanced DID method or the advanced regression method proposed in this study are useful tools in evaluating the policy effects of hierarchical data.

The key application of the DID model is to assess the mean effects of a policy intervention based on data from two time points. To evaluate the degree of linear change in policy effects over time or dose–response effects, panel data from multiple time points are required. For data from really long time series, fractional time series, DID models have been used.[20] In our case, data were available from five time points and so a series of four multilevel DID models had to be constructed with the common baseline year at 2008 and the second time point at 2009, 2010, 2011 and 2012, respectively, so that each DID model reflects a period of intervention of 1, 2, 3 and 4 years, respectively. Within the intervention period, facilities not yet exposed to the policy were classified as the control group. In this way, we were able to show the possible cumulative policy effects with the length of intervention time and consistent effects of covariates, as well as clustering effects at all three levels in the data structure of the four models. However, one can also choose to compare different time points, in which different control matches would be made, and different results could be observed for any effects of interest. Statistically, we cannot make reference to the possible dose–response effects based on four independent models but we can only provide a descriptive summary. We would also find it difficult to interpret results if other effects of covariates and random effects were different among the four models. In contrast, the multilevel RM model estimated the dose–response effects of the policy, the effects of covariates and random effects in one model, which is highly efficient. It is easy to make statistical inference with a clear interpretation.

In assessing the dose–response effects of the NEMP, time effects in a calendar year must be controlled for. To achieve this purpose, each multilevel DID model included a parameter $\beta_1$ to indicate the time effect in the particular year of intervention, and four models showed the year effects of 2009, 2010, 2011 and 2012, respectively, in comparison to the year 2008. The multilevel RM model treated time in a calendar year as a categorical covariate and estimated the year effects in contrast to 2008. It is obvious that the latter model handled the time effects with much statistical efficiency as shown by much smaller SEs of the year effects.

Although the trends of dose–response effects, the effects of covariates, and the distribution of components in the variation of the LG-OPV outcome estimated by the series of multilevel DID models and the multilevel RM model were similar, we observed differences in the SEs of estimates for those effects. For example, the SEs for the dose–response effects and the effect of time in a year were much smaller for the multilevel RM model than those for the multilevel DID models. The main reason is that each multilevel DID model only selected data between two time points and also lost some cases from the propensity matching process. In contrast, the multilevel RM model pooled data from all five time points into one joint model, which enlarges the sample size to almost five times the full dataset, hence there is much greater statistical efficiency in estimating time-related effects. This implies that when the sample size at each time point was small, the multilevel RM model was much more effective in detecting dose–response effects than the series of multilevel DID models.

Our study showed that the multilevel DID models and the multilevel RM model could break down variance of the outcome into components based on the level of data hierarchical structure, that is, calculate variances of random intercepts by levels. However, the former models cannot easily estimate variance in the dose–response effects for any of the levels in the dataset. Namely we obtained no answer to our research question (2) on whether such a policy effect was different among provinces or counties or facilities, and research question (3) on which could be the best or worst performers in the implementation of the policy. Based on the multilevel RM model we were able to show the variation estimates of the dose–response effects across the provinces, counties and facilities, respectively. We calculated the residuals to identify the highest and lowest dose–response effect among provinces for illustration purposes. We could do the same analysis to pinpoint counties and facilities using the same principle. The identified provinces or counties or facilities could be further examined to determine which context factors might be associated with the results of policy intervention and how to improve their situations.

Both models have shown that a higher ratio of health professional staff and higher assets index at the facility level were associated with higher usage of outpatient services, which agreed with a previous study.[20] The overall difference in the OPV or service use between rural and urban facilities and between the western and eastern regions in China in our study supported similar findings of a previous study.[20] However, these facility conditions only explained a small amount of variation in the dose–response effect of the policy intervention. Examination of other context factors such as social demographics, subculture environment and local policies are guaranteed.

This study has notable advantages. Earlier studies focusing on the effects of the NEMP were mainly cross-sectional; only two studies[20 34] by Gong and Li used the nationwide monitoring data. Gong's study included only 35 cities covering a 5-year period, from 2007 to 2011, and Li's study included the same data but using the conventional DID analysis. Using data from the annual national report of healthcare facilities, the results are representative nationwide and statistically robust due to

the large sample size and the fact that they are generalisable as a natural experiment in China. Although vast literature exists on the application of the conventional DID method to assess the overall effect of interventions, to the best of our knowledge this study is the first to illustrate advanced multilevel DID and multilevel RM models to analyse the dose–response effect based on hierarchically structured panel data, and to point to further analytic aspects of the multilevel RM model to investigate random effects attributable to potential local context effects in order to improve implementation of policy intervention with evidence of identified location or grassroots facilities.

The study has some limitations in the dataset. First, the exact starting time of NEMP implementation is different in different facilities, and we did not have data on the exact starting date but only the starting year. Interventions in certain facilities may not have had data for the whole year, which may lead to an underestimation of the dose–response effect of the NEMP on service use. Second, the study is not a real stepped wedge design, as the implementation of the NEMP at the facility level by time block was not randomised but by convenience.[39–41] However, we treated the time block representing the length of policy implementation as an independent variable in the multilevel RM models; lack of randomisation of such a variable will not affect the estimate of its effects. Third, the data were collected from the administrative health system, which might contain reporting errors and missing information at some time points. Despite these limitations, the study is methodology orientated, and comparison between models using the same dataset is not affected by those data issues. Instead the example analysis has provided important information on the dose–response effect of the NEMP in China and highlighted the importance of structural determinants of the NEMP effect, accounting for contextual factors in the future assessment of the NEMP effects.

## CONCLUSION

For hierarchically structured panel data, which are commonplace in national policy implementation, multilevel DID models and multilevel RM models should be employed to assess the dose–response effect of a policy. The latter model is statistically more efficient and easier to interpret than the former method and is a powerful tool to break down variation in the dose–response effects according to the level of hierarchical structure, thus identifying subjects with the best and worst results of effect for further investigation.

**Author affiliations**
[1]West China School of Public Health, Sichuan University, Chengdu, Sichuan, People's Republic of China
[2]West China Research Center for Rural Health Development, Sichuan University, Chengdu, Sichuan, People's Republic of China
[3]School of Medicine, University of Nottingham, Nottingham, UK
[4]Swinburne University of Technology, Victoria, Australia
[5]Center for Health Statistical Information, National Health and Family Planning Commission of the People's Republic of China, Beijing, People's Republic of China

**Acknowledgements** The authors thank all PHF facility workers who submitted the data and the Center for Information and Statistics, which collected the data.

**Contributors** MY and YR designed the study, performed data analysis, interpreted results and drafted the manuscript. QL, FC and JP contributed to the original design for data collection and data cleaning, made critical comments on data analysis and helped draft the manuscript. XL was responsible for the study that generated the subset data for the current study. QM was the PI of the original study, which collected data for subsequent secondary data analysis. All authors read and approved the final version of this manuscript.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.



## REFERENCES

1. Abadie A. Semiparametric difference-in-differences estimators. *Rev Econ Stud* 2005;72:1–19.
2. Ikenwilo D. A difference-in-differences analysis of the effect of free dental check-ups in Scotland. *Soc Sci Med* 2013;83:10–8.
3. Jing S, Yin A, Shi L, *et al*. Whether new cooperative medical schemes reduce the economic burden of chronic disease in rural China. *PLoS ONE* 2013;8:e53062.
4. Chen Y, Jin GZ. Does health insurance coverage lead to better health and educational outcomes? Evidence from rural China. *J Health Econ* 2012;31:1–14.
5. Nelson RE, Hicken B, West A, *et al*. The effect of increased travel reimbursement rates on health care utilization in the VA. *J Rural Health* 2012;28:192–201.
6. Wagner AK, Soumerai SB, Zhang F, *et al*. Segmented regression analysis of interrupted time series studies in medication use research. *J Clin Pharm Ther* 2002;27:299–309.
7. Wagstaff A, Lindelow M, Jun G, *et al*. Extending health insurance to the rural population: an impact evaluation of China's new cooperative medical scheme. *J Health Econ* 2009;28:1–19.
8. Li L, Liang LJ, Wu Z, *et al*. Institutional support for HIV/AIDS care in China: a multilevel analysis. *AIDS Care* 2008;20:1190–6.
9. Grytten J, Holst D, Skau I. Per capita remuneration of dentists and the quality of dental services. *Community Dent Oral Epidemiol* 2013;41:395–400.
10. Arrieta A. The impact of the Massachusetts health care reform on unpaid medical bills. *Inquiry* 2013;50:165–76.
11. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006;6:54.
12. Mdege ND, Man MS, Taylor N *et al*. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;64:936–48.
13. Hemming K, Haines TP, Chilton PJ, *et al*. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;350:h391.
14. Murphy DA, Harrell L, Fintzy R, *et al*. A comparison of methamphetamine users to a matched NHANES cohort: propensity score analyses for oral health care and dental service need. *J Behav Health Ser Res* 2016;43:676–90.
15. The State Council of China. Implementation plan for the recent priorities of the health care system reform (2009–2011). 2009.

16. You X, Kobayashi Y. The new cooperative medical scheme in China. *Health Policy* 2009;91:1–9.
17. Ouyang Y. China relaxes its one-child policy. *Lancet* 2013;382:e28.
18. Yang DP, Chai CQ, Zhu YN, *et al*. *The China educational development yearbook*. Leiden: Brill:13ff. ISBN 90–04–17178-9;2009.
19. Fang Y, Wagner AK, Yang S, *et al*. Access to affordable medicines after health reform: evidence from two cross-sectional surveys in Shaanxi Province, western China. *Lancet Glob Health* 2013;1:e227–37.
20. Li Q. *Policy evaluation of China National Essential Medicines: a difference-in-difference analysis with propensity score matching*. [PhD dissertation]. Sichuan University, 2014: Y2014/R01/012.
21. *China statistical yearbook*; National Bureau of Statistics. 2013.
22. Angrist JD, Pischke J. *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press, 2008.
23. Heimeshoff M, Schreyogg J, Tiemann O. Employment effects of hospital privatization in Germany. *Eur J Health Econ* 2014;15:747–57.
24. Coca-Perraillon M, *Matching with propensity scores to reduce bias in observational studies*. Burlington, MA: Adheris Inc., http://www.lexjansen.com/nesug/nesug06/an/da13.pdf.2006
25. Parsons LS. *Reducing bias in a propensity score matched-pair sample using greedy matching techniques*. Seattle, WA: Ovation Research Group, http://www2.sas.com/proceedings/sugi26/p214-26.pdf
26. Myers ND, Brincks AM, Ames AJ, *et al*. Multilevel modeling in psychosomatic medicine research. *Psychosom Med* 2012;74:925–36.
27. Karlsdotter K, Martin Martin JJ, Lopez del Amo Gonzalez MP. Multilevel analysis of income, income inequalities and health in Spain. *Soc Sci Med* 2012;74:1099–106.
28. Rasbash J, Browne W, Goldstein H, *et al*. *A user's guide to MLwiN*. London: Institute of Education, University of London, 2000.
29. Cameron A, Ewen M, Ross-Degnan D, *et al*. Medicine prices, availability, and affordability in 36 developing and middle-income countries: a secondary analysis. *Lancet* 2009;373:240–9.
30. Song Y, Bian Y, Petzold M, *et al*. Effects of The National Essential Medicine System in reducing drug prices: an empirical study in four Chinese provinces. *J Pharm Policy Pract* 2014;7:12.
31. Yang H, Dib HH, Zhu MM, *et al*. Prices, availability and affordability of essential medicines in rural areas of Hubei Province, China. *Health Policy Plan* 2010;25:219–29.
32. Zhang WY, Li YR, Li YJ, *et al*. A cross-sectional analysis of prescription and stakeholder surveys following essential medicine reform in Guangdong Province, China. *BMC Health Serv Res* 2015;15:98.
33. Li Y, Ying C, Sufang, G, *et al*. Evaluation, in three provinces, of the introduction and impact of China's National Essential Medicines Scheme. *Bull World Health Organ* 2013;91:184–94.
34. Gong Y, Yang C, Yin X, *et al*. The effect of essential medicines programme on rational use of medicines in China. *Health Policy Plann* 2016;31:21–7.
35. Chen MS, Wang LJ, Chen W, *et al*. Does economic incentive matter for rational use of medicine? China's Experience from the Essential Medicines Program. *Pharmacoeconomics* 2014;32:245–55.
36. Xu SM, Bian C, Wang H, *et al*. Evaluation of the implementation outcomes of the Essential Medicines System in Anhui county-level public hospitals: a before-and-after study. *BMC Health Serv Res* 2015;15:403.
37. Goldstein H. *Multilevel Statistical Models*. 4th edn. London: Wiley, 2011.
38. Voyer D, Voyer SD, Tramonte L. Free-viewing laterality tasks: a multilevel meta-analysis. *Neuropsychology* 2012;26:551–67.
39. Beard E, Lewis JJ, Copas A, *et al*. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015;16:353.
40. Copas AJ, Lewis JJ, Thompson JA, *et al*. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials* 2015;16:352.
41. Davey C, Hargreaves J, Thompson JA, *et al*. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials* 2015;16:358.