

RESEARCH ARTICLE

# Application of a Novel Grey Self-Memory Coupling Model to Forecast the Incidence Rates of Two Notifiable Diseases in China: Dysentery and Gonorrhoea

Xiaojun Guo<sup>1,2\*</sup>, Sifeng Liu<sup>2</sup>, Lifeng Wu<sup>2</sup>, Lingling Tang<sup>3</sup>

1. School of Science, Nantong University, Nantong, Jiangsu Province, China, 2. College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu Province, China, 3. School of Electrical and Computer Engineering, Cornell University, Ithaca, New York, United States of America

\*[guoxj159@163.com](mailto:guoxj159@163.com)



 OPEN ACCESS

**Citation:** Guo X, Liu S, Wu L, Tang L (2014) Application of a Novel Grey Self-Memory Coupling Model to Forecast the Incidence Rates of Two Notifiable Diseases in China: Dysentery and Gonorrhoea. PLoS ONE 9(12): e115664. doi:10.1371/journal.pone.0115664

**Editor:** Hiroshi Nishiura, The University of Tokyo, Japan

**Received:** May 28, 2014

**Accepted:** November 22, 2014

**Published:** December 29, 2014

**Copyright:** © 2014 Guo et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability:** The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper.

**Funding:** This work was partially supported by the National Natural Science Foundation of China (No. 71401051) and the Funding of Nantong Science and Technology Program (No. HS2013026). But the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

## Abstract

**Objective:** In this study, a novel grey self-memory coupling model was developed to forecast the incidence rates of two notifiable infectious diseases (dysentery and gonorrhoea); the effectiveness and applicability of this model was assessed based on its ability to predict the epidemiological trend of infectious diseases in China.

**Methods:** The linear model, the conventional GM(1,1) model and the GM(1,1) model with self-memory principle (SMGM(1,1) model) were used to predict the incidence rates of the two notifiable infectious diseases based on statistical incidence data. Both simulation accuracy and prediction accuracy were assessed to compare the predictive performances of the three models. The best-fit model was applied to predict future incidence rates.

**Results:** Simulation results show that the SMGM(1,1) model can take full advantage of the systematic multi-time historical data and possesses superior predictive performance compared with the linear model and the conventional GM(1,1) model. By applying the novel SMGM(1,1) model, we obtained the possible incidence rates of the two representative notifiable infectious diseases in China.

**Conclusion:** The disadvantages of the conventional grey prediction model, such as sensitivity to initial value, can be overcome by the self-memory principle. The novel grey self-memory coupling model can predict the incidence rates of infectious diseases more accurately than the conventional model, and may provide useful references for making decisions involving infectious disease prevention and control.

## Introduction

Prevention and control of infectious diseases is an increasingly important public health issue. After World War II, developments in preventive medicine, clinical medicine, and preclinical medicine have served major roles in preventing and controlling infectious diseases. Many acute or chronic infectious diseases have been effectively controlled by the use of antibiotics and vaccines [1, 2]. However, the chronicity, arduousness, and complexity of infectious disease prevention and control are often ignored, which has resulted in the emergence of new pathogens with changes in the environment and the expansion of domestic and international communication [3, 4]. The World Health Organization has declared that epidemic factors spreading infectious diseases include global population movement, emergence of drug-fast pathogenic microorganisms, global climatic variation, social unstable factors, and flaws in health management [5]. Strategies to prevent and control infectious diseases are particularly urgent in developing countries with a weak economy, poor sanitary condition, large population density, and significant international and domestic population movement [6, 7]. Previous experiences have shown that accurate forecasts and analyses of the future trend of infectious diseases can realize timely epidemic detection and prevention. Establishing long-term prevention strategies can lay the foundation for early warning of diseases and provide a theoretical basis for drafting prevention strategies and solutions. Therefore, early warning and forecast of infectious diseases is important for health care and public hygiene management.

A statistical analytical method that combines mathematics and infectious disease epidemiology is used to predict the emergence of infectious diseases. This method has been widely applied for early warning and forecast of all types of infectious diseases. Model-based prediction methods for epidemics have been a major research focus in public health. Studies commonly combine statistical analysis (i.e., regression analysis [8] and time series analysis [9]) and mathematical modeling (i.e., epidemic dynamic model [10, 11]) to forecast epidemic regularity and development trends based on data related to infectious diseases. These methods have been actively researched for their utility in generating early warnings against disease outbreaks [12, 13]. In particular, prediction of disease incidence rates is a popular topic of research. Disease forecasts can provide important references for public hygiene management, and serve as effective resources for planning, prevention and control. With the development of powerful prediction theories, statistical regression models [14–16], time-sequences models [17, 18], discrete time stochastic model [19], age-structured epidemic model [20], grey system models [21, 22], Markov chain models [23, 24], and artificial neural network models [25, 26] have been applied to predict future epidemiological trends. However, common statistical prediction methods have limitations. For example, large samples of historical data that follow a certain representative statistical distribution are required. The evolutionary path of an infectious disease is influenced by uncertainties and characterized as a grey system. Hence, the occurrence and prevalence of an infectious disease can be regarded as a

typical dynamic variable grey system. Importantly, grey system models are practical because data sample size and probability distribution are not strictly required.

When dealing with small samples and incomplete information, traditional methods (i.e., probability statistics and fuzzy mathematics) show limitations [27, 28] such as requirement of a large amount of statistical data and a structuralized system. To overcome these disadvantages, Deng first proposed the grey systems theory in 1982 to study the uncertainty of systems with unclear statistical distribution [29]. The grey model uses the accumulated generating operation to process raw data to reduce their degree of stochasticity and to increase their regularity [30, 31]. As the simplest model, the GM(1,1) model (univariate first-order differential equation) is especially appropriate for predicting the overall development tendency of the dynamical system (i.e., approximate exponential increasing tendency and exponential decreasing tendency). It has been broadly applied in numerous fields, such as economy and management, industry and agriculture, medicine and health, and engineering sciences [32–37]. However, the conventional GM(1,1) model essentially belongs to the initial value solving problem of differential equations which only meet the initial condition at one point, i.e., the observed values at one moment. Accordingly, the original dynamical differential equation has the limitation of being sensitive to initial values, and that becomes a disadvantage when historical information is not fully available. Recently, many scholars have focused on improving the prediction performance of the grey prediction models. The self-memory principle is one of the most important methods to enhance the precision of a model.

On the basis of retrieved modeling methods, the self-memory principle of dynamic system was first proposed by Cao in 1993 [38]. As a statistically dynamic method to solve problems of nonlinear dynamic systems, it successfully integrated determinism and random theories with mathematics [39]. The self-memory principle can retrieve ideal nonlinear dynamic models from practical observational data. It overcomes the weakness of being sensitive to initial values for differential equations, as well as the limitation of irrelevance to the mechanism modeling due to utilization of historical materials. The method is a breakthrough for numerical solution of traditional initial-value problems and statistical approaches. This novel prediction model combines the advantages of the self-memory principle and the grey GM(1,1) model by coupling their prediction methods. Its excellent predictive performance lies in the fact that the weakness of conventional GM(1,1) model, i.e., sensitivity to initial value, can be overcome by using a multi-time-point initial field instead of a single-time-point initial field. The concept has been utilized increasingly in time series forecasting in multiple fields, such as meteorology, engineering, and economics [40, 41]. In recent years, some scholars have attempted to introduce the self-memory principle into certain basic grey prediction models [42–44]. However, these published research methods have been applied mainly in the fields of meteorology, hydrography, and engineering science, and only minimally in medicine and public health.

In the fields of epidemiology and public hygiene, the occurrence and prevalence of an infectious disease are commonly accompanied by irregular individual fluctuations due to many internal and external unstable factors. Hence, the spread and prevalence of an infectious disease can be regarded as a dynamical variable grey system with stochastic fluctuation. The grey prediction models have been effectively utilized for forecast and analysis of the morbidity and mortality of epidemics, such as parasitosis and phymatosis [21, 37]. In this paper, in light of the uncertainty features of infectious diseases, two representative infectious diseases (dysentery and gonorrhoea) among categories A and B of notifiable diseases in China were selected to predict their incidence rates by coupling the GM(1,1) model and the self-memory principle. Although most infectious diseases have been successfully controlled in China, handling of epidemics of some infectious diseases in certain regions are not going well [22]. Dysentery is a major public health issue in many countries in the world [45]. Transmission of dysentery is fecal-oral, which may involve polluted food, water, daily contact, and flies. Despite the fact that the incidence of intestinal infectious diseases has declined considerably in recent years worldwide, the incidence of dysentery remains high in developing countries. Dysentery is one of the most common epidemics in overcrowded areas with inadequate sanitation. Notably, dysentery is a recurrent challenge in many parts of the world. In addition, gonorrhoea, a bacterial infection caused by *Neisseria gonorrhoeae*, is a highly communicable sexually transmitted infection and, due to its short incubation period, may serve as an indicator of recent risky sexual behavior in symptomatic cases [46]. Furthermore, in developing countries, sexually transmitted diseases and their complications are among the top five conditions for which adults seek medical care. These conditions may cause acute or chronic symptoms as well as delayed sequels such as infertility, ectopic pregnancy, cervical cancer and premature fatalities among infants and adults. Consequently, dysentery and gonorrhoea rank top in categories A and B of notifiable diseases in China.

In this study, prediction performances of the SMGM(1,1) model, the conventional GM(1,1) model and the linear model were compared. The model with the best fit was then utilized to deduce emerging epidemic tendencies. We propose that the grey self-memory coupling prediction model is appropriate for forecasting the incidence rates of infectious diseases in China. Results of our analyses may provide effective guidance in the decision-making process for the prevention and control of epidemics.

## Materials and Methods

### Data sources

In this study, the incidence rates of two types of notifiable infectious diseases (dysentery and gonorrhoea) were investigated. The incidence rates of the two infectious diseases were obtained from public governmental statistical data, published by the China Health Statistical Yearbook of 2013 [47]. Since the

Chinese Centre for Disease Control and Prevention was founded in 2002, and it experienced the ordeal of the SARS epidemic the following year, only statistical data from 2004 to 2012 were selected for modeling analysis.

## Methods

The linear model was formulated as  $y=mx+b$ . Detailed principles of modeling analysis and accuracy assessment of the grey self-memory coupling prediction method have been described previously by Wu et al. [31] and Cao et al. [48]. Excel software, Grey system software 3.0 and MATLAB software 7.0 were utilized for modeling and simulation.

### A. The principle of grey self-memory coupling prediction

After the accumulated original statistical data are generated and the moving average is calculated, the grey system GM(1,1) model can weaken the randomness of the original data and generate regular cumulative data. Consequently, the prediction model can be approximated by the solution of a linear first-order differential equation. Nevertheless, the self-memory principle not only emphasizes the overall exponential development tendency of the dynamical system, but also its individual stochastic fluctuations. After introducing the memory function, which contains historical information, into the system's dynamic differential equation, the equation can be transformed into a difference-integral equation, called self-memorization equation, by defining the inner product in Hilbert space. Because the systematic self-memorization equation contains multiple time-point initial fields instead of only single time-point initial field, the weakness of being sensitive to the initial value of the original dynamic differential equation can be overcome. By studying systematic inner memorability, the systematic overall exponential development tendency with individual stochastic fluctuations can be modeled and predicted. The superiority of the self-memory principle lies in the fact that the systematic predictive ability can be improved by not only combining dynamics calculations and estimating parameters of historical data, but also extracting systematic information from historical data in statistics.

### B. Model construction

For our modeling analysis, the incidence data of the two infectious diseases from 2004 to 2011 were taken as the modeling samples (i.e., original series  $X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$ ). Furthermore, the incidence data of the year 2012 were selected as the testing samples for the prediction test.

First, the general procedure for a conventional GM(1,1) model was derived as follows:

Step 1. Assume that the sequence  $X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$  is an original non-negative data sequence, where  $x^{(0)}(k)$  is the time series data at time  $k$ . The sequence  $X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n)\}$  is the first-order accumulated generation sequence of  $X^{(0)}$ , where  $x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i)$ ,  $k = 1, 2, \dots, n$ .

Step 2. The basic form of the GM(1,1) model is defined as

$$x^{(0)}(k) + az^{(1)}(k) = b, (z^{(1)}(k) = \frac{x^{(1)}(k) + x^{(1)}(k+1)}{2})$$

where the parameters  $a$  and  $b$  are called developing and grey input coefficients, respectively. Let sampling time  $\Delta t = 1$ ; then, by the least square method, the parameters  $a$  and  $b$  can be obtained as

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (B^T B)^{-1} B^T X_n$$

where

$$X_n = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}, B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix}$$

Step 3. The whitening equation of the GM(1,1) model is given by  $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$ . By making the initial value  $\hat{x}^{(1)}(1) = x^{(0)}(1)$ , the time response sequence of the GM(1,1) model is given by

$$\hat{x}^{(1)}(k+1) = (x^{(0)}(1) - \frac{b}{a})e^{-ak} + \frac{b}{a} \tag{1}$$

and the simulative value of sequence  $X^{(1)}$  can be obtained from **Eq. 1** accordingly. Consider the inverse accumulated generation

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) = (1 - e^a)(x^{(0)}(1) - \frac{b}{a})e^{-ak}, k = 1, 2, \dots, n-1$$

Thus, the simulative value of sequence  $X^{(0)}$  can be obtained. The procedure followed to develop a novel SMGM (1,1) model is as follows:

Step 1. Determining the self-memory dynamic equation.

Let  $dx^{(1)}/dt$  in the whitening differential equation of the GM(1,1) model be  $F(x,t)$ . Then,

$$F(x,t) = -ax^{(1)} + b \tag{2}$$

The differential equation  $dx^{(1)}/dt$ , determined by **Eq. 2**, is considered systematic self-memory dynamic equation of the grey self-memory coupling model:

$$\frac{dx}{dt} = F(x,t) \tag{3}$$

where  $x$  is a variable,  $t$  is time interval series, and  $F(x,t)$  is the dynamic kernel. Then, a memory function  $\beta(t)$  was introduced and an inner product in the Hilbert space was defined:

$$(f,g) = \int_{a_0}^{b_0} f(\xi)g(\xi)d\xi \quad (f,g \in L^2) \tag{4}$$

Step 2. Deducing the difference-integral equation.

Let one time set  $T = [t_{-p}, t_{-p+1}, \dots, t_{-1}, t_0, t]$ , where  $t_{-p}, t_{-p+1}, \dots, t_{-1}, t_0$  is historical observation time,  $t_0$  is predicted initial time,  $t$  is coming prediction time, the retrospective order of the equation is  $p$  and time sampling interval is  $\Delta t$ . After applying the above inner product operation into **Eq. 3** and supposing that variables  $x, \beta$  are continuous, differentiable, and integrable, the analytic formula of **Eq. 3** is obtained as  $\int_{t_{-p}}^t \beta(\tau) \frac{\partial x}{\partial \tau} d\tau = \int_{t_{-p}}^t \beta(\tau) F(x, \tau) d\tau$ , that is,

$$\int_{t_{-p}}^{t_{-p+1}} \beta(\tau) \frac{\partial x}{\partial \tau} d\tau + \int_{t_{-p+1}}^{t_{-p+2}} \beta(\tau) \frac{\partial x}{\partial \tau} d\tau + \dots + \int_{t_0}^t \beta(\tau) \frac{\partial x}{\partial \tau} d\tau = \int_{t_{-p}}^t \beta(\tau) F(x, \tau) d\tau \tag{5}$$

For every integral term in the left-hand side of **Eq. 5**, after applying calculus and performing integration by parts, applying the median theorem, and performing algebraic operations, a difference-integral equation is deduced as:

$$\beta_t x_t - \beta_{-p} x_{-p} - \sum_{i=-p}^0 x_i^m (\beta_{i+1} - \beta_i) - \int_{t_{-p}}^t \beta(\tau) F(x, \tau) d\tau = 0 \tag{6}$$

where  $\beta_t \equiv \beta(t)$ ,  $x_t \equiv x(t)$ ,  $\beta_i \equiv \beta(t_i)$ ,  $x_i \equiv x(t_i)$ ,  $i = -p, -p+1, \dots, 0$ , and mid-value  $x_i^m \equiv x(t_m)$ ,  $t_i < t_m < t_{i+1}$ .

Step 3. Discretizing the self-memory prediction equation.

Let  $x_{-p} \equiv x_{-p-1}^m$  and  $\beta_{-p-1} \equiv 0$ , **Eq. 6** can be converted into

$$x_t = \frac{1}{\beta_t} \sum_{i=-p-1}^0 x_i^m (\beta_{i+1} - \beta_i) + \frac{1}{\beta_t} \int_{t_{-p}}^t \beta(\tau) F(x, \tau) d\tau = S_1 + S_2 \tag{7}$$

which is called self-memory equation with the retrospective order  $p$ . The first term  $S_1$ , defined as the self-memory term in **Eq. 7**, denotes the relative contributions of historical data at  $p+1$  times to the value of variable  $x_t$ . The second term  $S_2$ , defined as the exogenous effect term, is the total contribution of the function  $F(x,t)$  in the retrospective time interval  $[t_{-p}, t_0]$ . **Eq. 7** emphasizes serial correlation of the system by itself, i.e., the self-memory characteristic of the system. Therefore, **Eq. 7** is the self-memory prediction equation of the system. If integral operation is substituted by summation and differential is transformed into difference in **Eq. 7**, then the mid-value  $x_i^m$  is replaced by two values of different times, namely,



$x_i^m = \frac{1}{2}(x_{i+1} + x_i) \equiv y_i$ . By taking equidistance time interval  $\Delta t_i = t_{i+1} - t_i = 1$ , and merging  $\beta_t$  and  $\beta_i$  together, the self-memory equation of discrete form is shown as follows:

$$x_t = \sum_{i=-p-1}^{-1} \alpha_i y_i + \sum_{i=-p}^0 \theta_i F(x, i) \tag{8}$$

where  $\alpha_i = (\beta_{i+1} - \beta_i) / \beta_t, \theta_i = \beta_i / \beta_t$ .  $\alpha_i$  and  $\theta_i$  are called memory coefficients, and  $F(x, t)$  is determined by the dynamic kernel  $-ax^{(1)} + b$  of the GM(1,1) model.

Step 4. Solving the self-memory prediction model.

Assuming that there are  $L$  items of historical data, the memory coefficients  $\alpha_i$  and  $\theta_i$  can be estimated by the least square method. Let

$$X_t = \begin{bmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tL} \end{bmatrix}_{L \times 1}, \quad Y = \begin{bmatrix} y_{-p-1,1} & y_{-p,1} & \cdots & y_{-1,1} \\ y_{-p-1,2} & y_{-p,2} & \cdots & y_{-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{-p-1,L} & y_{-p,L} & \cdots & y_{-1,L} \end{bmatrix}_{L \times (p+1)}, \quad A = \begin{bmatrix} \alpha_{-p-1} \\ \alpha_{-p} \\ \vdots \\ \alpha_{-1} \end{bmatrix}_{(p+1) \times 1}$$

$$\Gamma = \begin{bmatrix} F(x, -p)_1 & F(x, -p+1)_1 & \cdots & F(x, 0)_1 \\ F(x, -p)_2 & F(x, -p+1)_2 & \cdots & F(x, 0)_2 \\ \vdots & \vdots & \ddots & \vdots \\ F(x, -p)_L & F(x, -p+1)_L & \cdots & F(x, 0)_L \end{bmatrix}_{L \times (p+1)}, \quad \Theta = \begin{bmatrix} \theta_{-p} \\ \theta_{-p+1} \\ \vdots \\ \theta_0 \end{bmatrix}_{(p+1) \times 1}$$

Then, Eq. 8 can be expressed in matrix form as follows:

$$X_t = YA + \Gamma\Theta \tag{9}$$

Let  $Z = [Y, \Gamma]$ ,  $W = \begin{bmatrix} A \\ \Theta \end{bmatrix}$ , then Eq. 9 turns into  $X_t = ZW$ , thereby  $W$  is obtained by the least square method:  $W = (Z^T Z)^{-1} Z^T X_t$ . When the memory coefficients matrix  $W$  is obtained, the simulation and prediction of original data sequence  $X^{(0)}$  can be performed. For the simulated and predicted value  $\hat{x}^{(1)}(t)$  of the accumulated generation sequence in grey self-memory coupling model, its inverse accumulated value  $\hat{x}^{(0)}(t)$  can be obtained as follows:

$$\hat{x}^{(0)}(t+1) = \hat{x}^{(1)}(t+1) - \hat{x}^{(1)}(t)$$

where  $t = 1, 2, \dots, n-1$  and  $\hat{x}^{(1)}(1) = x^{(0)}(1)$ .

### C. Modeling simulation and prediction accuracy assessment

Simulation and prediction accuracy is an important criterion for evaluating prediction models. Accuracy test must be performed to evaluate the rationality and reliability of prediction models before extrapolation and application.



**Table 1.** List of variance ratio and small error probability obtained in the accuracy test.

Modeling accuracy class	Test index	
	Variance ratio C	Small error probability p
1st level (superior)	≤ 0.35	≥ 0.95
2nd level (qualified)	0.35 ~ 0.50	0.80 ~ 0.95
3rd level (marginal)	0.50 ~ 0.65	0.70 ~ 0.80
4th level (disqualified)	≥ 0.65	≤ 0.70

doi:10.1371/journal.pone.0115664.t001

Consequently, two popular test criteria such as variance ratio and small error probability [22] were used to compare the accuracy of different prediction models, as shown in Table 1.

$\sigma_1 = \sqrt{\frac{1}{n} \sum_{k=1}^n (x^{(0)}(k) - \bar{x})^2}$  and  $\sigma_2 = \sqrt{\frac{1}{n} \sum_{k=1}^n (\varepsilon(k) - \bar{\varepsilon})^2}$  are the mean square error of original values and residual error, respectively. Given  $C_0 > 0$ , if the variance ratio  $C = \sigma_2 / \sigma_1 < C_0$ , then the model is considered to pass through the variance ratio test [22].

In the same way, given  $p_0 > 0$ , if the small error probability  $p = P(|\varepsilon(k) - \bar{\varepsilon}| < 0.6745\sigma_1) > p_0$ , then the model is supposed to pass through the small error probability test [22].

The absolute percentage error at time  $k$  is denoted by

$$APE(k) = \left| \frac{\hat{x}^{(0)}(k) - x^{(0)}(k)}{x^{(0)}(k)} \right| \times 100\%,$$

and the mean absolute percentage error at all

times is defined as  $MAPE = \frac{1}{n-1} \sum_{k=2}^n APE(k)$ . Accordingly, comparisons between the actual values and simulative values derived from each prediction models can be analyzed using APE(k) and MAPE.

#### D. Programming procedure of MATLAB software

The calculation was performed, as mentioned above, with the help of MATLAB software for computational efficiency. The programming procedure for the SMGM(1,1) model is shown in Fig. 1.

#### Data analysis

The incidence rates of two representative infectious diseases (dysentery and gonorrhoea) from 2004 to 2012 in China were adopted to demonstrate the practicability and effectiveness of the proposed SMGM(1,1) model and its superiority to the linear model and the conventional GM(1,1) model. On the whole, the time series of incidence rates of dysentery and gonorrhoea all showed an

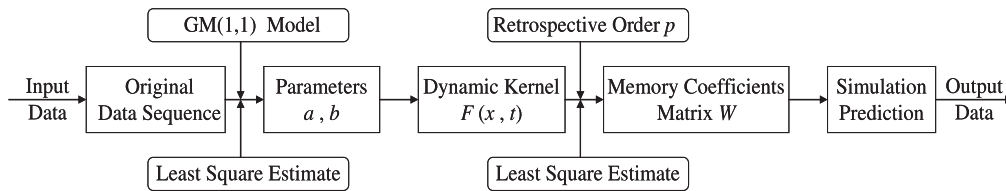


Fig. 1. Programming procedure for the SMGM(1,1) model.

doi:10.1371/journal.pone.0115664.g001

obvious exponential decreasing tendency. They were accompanied by some irregular individual fluctuations due to unstable changes imposed by social and economic factors. Therefore, two SMGM(1,1) models were established to model and predict the incidence rates, and were compared with their corresponding linear models and conventional GM(1,1) models. When performing the modeling analysis, the incidence rates of the first eight time-points were taken as modeling samples, and the data of the ninth time-point was selected as the test sample for the prediction test. In accordance with the principles and the steps mentioned above, two SMGM(1,1) models were established to forecast incidence rates. The value of retrospective order was uniformly determined as  $p = 1$  by trial calculation under the principle of minimum error. APE(k) and MAPE were used to compare the actual values with the simulated values to evaluate the predictive performance compared to linear models and conventional GM(1,1) models.

## Results

### Forecasting the incidence rates of dysentery

Based on the incidence data of dysentery from 2004 to 2011, the differential equation of conventional GM(1,1) model can be formulated as follows:

$$\frac{dx^{(1)}}{dt} = -0.1250x^{(1)} + 42.4204 \quad (10)$$

Let the right-side term of Eq. 10 be the dynamic kernel  $F(x,t)$ . Then, the systematic self-memory dynamic equation  $dx/dt = F(x,t)$  of SMGM(1,1) model is obtained. After applying the inner product operation (4) into  $dx/dt = F(x,t)$ ,

then the analytic formula is obtained as  $\int_{t-1}^{t_0} \beta(\tau) \frac{\partial x}{\partial \tau} d\tau + \int_{t_0}^t \beta(\tau) \frac{\partial x}{\partial \tau} d\tau =$

$\int_{t-1}^t \beta(\tau) F(x,\tau) d\tau$ . According to the modeling steps mentioned above, a difference-integral equation is deduced as

$$\beta_t x_t - \beta_{-1} x_{-1} - \sum_{i=-1}^0 x_i^m (\beta_{i+1} - \beta_i) - \int_{t-1}^t \beta(\tau) F(x,\tau) d\tau = 0, \text{ and the self-memory}$$

$$\text{prediction equation is obtained as } x_t = \frac{1}{\beta_t} \sum_{i=-2}^0 x_i^m (\beta_{i+1} - \beta_i) +$$

$\frac{1}{\beta_t} \int_{t-1}^t \beta(\tau)F(x, \tau) d\tau$ . After the integral operation is substituted by summation and the differential is transformed into difference, the discrete form of self-memory equation for dysentery incidence can be expressed as

$x_t = \sum_{i=-2}^{-1} \alpha_i y_i + \sum_{i=-1}^0 \theta_i F(x, i)$ , where  $\alpha_i = (\beta_{i+1} - \beta_i) / \beta_t$ ,  $\theta_i = \beta_i / \beta_t$ . Using the least square estimation method, the memory coefficients matrix can be obtained as

$$W = [\alpha_{-2} \quad \alpha_{-1} \quad \theta_{-1} \quad \theta_0]^T = [0.1784 \quad 0.8344 \quad 12.4231 \quad -12.4821]^T$$

The simulated values and errors of the linear model, the conventional GM(1,1) model and the SMGM(1,1) model are presented in [Table 2](#). The variance ratio and small error probability of the three models are all at the first level, as shown in [Table 1](#). Since the three models passed the simulation accuracy assessment, they could be used to perform predictions. From the APE(k) and MAPE of fitting values, as shown in [Table 2](#), the simulated precision of the novel SMGM(1,1) model is markedly superior to that of the other two models. Considering the incidence rate in 2012, the SMGM(1,1) model also exhibits better single-step predictive performance compared with the other models. Using the SMGM(1,1) model, the next incidence rate of dysentery in China is predicted to be 14.10 per 100,000.

### Forecasting the incidence rates of gonorrhoea

Based on the incidence data of gonorrhoea from 2004 to 2011, the differential equation of the conventional GM(1,1) model is formulated as follows:

$$\frac{dx^{(1)}}{dt} = -0.1065x^{(1)} + 16.2546$$

Then, the prediction equation of gonorrhoea incidence can be similarly obtained by using the same formula above, where the memory coefficients matrix is

$$W = [\alpha_{-2} \quad \alpha_{-1} \quad \theta_{-1} \quad \theta_0]^T = [0.0019 \quad 0.9989 \quad -1.7813 \quad 3.4540]^T$$

The simulated values and errors of the linear model, the conventional GM(1,1) model and the SMGM(1,1) model are presented in [Table 3](#). All models passed the simulation accuracy assessment, and the simulated and single-step predictive precisions of the SMGM(1,1) model were also superior to other two models. Thus, the next incidence rate of gonorrhoea in China is estimated at 6.05 per 100,000 based on the SMGM(1,1) model.

Furthermore, [Figs. 2–5](#) illustrate the fitting results of the simulated curves obtained by the three compared models with the original incidence curves of dysentery and gonorrhoea, and their corresponding comparison results of relative

**Table 2.** Simulated values and errors of the linear model, the GM(1,1) model and the SMGM(1,1) model for the incidence rate of dysentery ( $1/10^5$ ).

Year	Original Value	Linear model		GM(1,1) model		SMGM(1,1) model	
		Simulative value	APE	Simulative value	APE	Simulative value	APE
2004	38.30	37.91	1.01%	38.3	—	—	—
2005	34.92	34.77	0.44%	35.38	1.32%	—	—
2006	32.36	31.62	2.28%	31.22	3.52%	32.40	0.12%
2007	27.99	28.47	1.73%	27.55	1.57%	27.45	1.93%
2008	23.43	25.33	8.10%	24.32	3.80%	24.60	4.99%
2009	20.45	22.18	8.47%	21.46	4.94%	19.93	2.54%
2010	18.90	19.03	0.71%	18.94	0.21%	18.42	2.54%
2011	17.74	15.89	10.44%	16.71	5.81%	18.06	1.80%
MAPE			4.15%		3.02%		2.32%
2012	15.40	12.74	17.26%	14.75	4.22%	15.12	1.82%

There is no simulated value for the first two time-points because of the retrospective order  $p=1$ .

doi:10.1371/journal.pone.0115664.t002

percentage error distribution. From the comparative analysis, the prediction accuracy of the proposed SMGM(1,1) model is remarkably higher than that of the linear model and the conventional GM(1,1) model. Consequently, the proposed SMGM(1,1) model can better catch the tendency of integral development and individual variation of original data, and is a reliable and stable prediction model for predicting the future development tendency of infectious diseases.

### Discussion

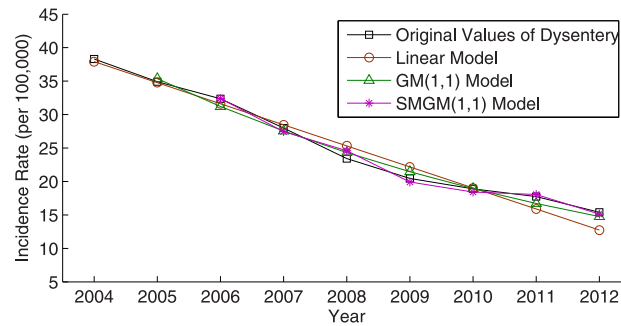
The GM(1,1) model, as a basic and typical grey system prediction model, mainly reflects the statistical laws of diseases by processing the corresponding statistical

**Table 3.** Simulated values and errors of the linear model, the GM(1,1) model and the SMGM(1,1) model for the incidence rate of gonorrhoea ( $1/10^5$ ).

Year	Original Value	Linear model		GM(1,1) model		SMGM(1,1) model	
		Simulative value	APE	Simulative value	APE	Simulative value	APE
2004	17.34	15.30	11.78%	17.34	—	—	—
2005	13.79	14.10	2.22%	13.67	0.87%	—	—
2006	12.14	12.90	6.23%	12.29	1.24%	12.14	0.00%
2007	11.08	11.70	5.55%	11.05	0.27%	11.11	0.27%
2008	9.90	10.49	6.00%	9.93	0.30%	9.85	0.51%
2009	9.02	9.29	3.03%	8.93	1.00%	8.96	0.67%
2010	7.91	8.09	2.31%	8.03	1.52%	8.03	1.52%
2011	7.31	6.89	5.71%	7.21	1.37%	7.28	0.41%
MAPE			5.36%		0.94%		0.56%
2012	6.82	5.69	16.55%	6.49	4.84%	6.64	2.64%

There is no simulated value for the first two time-points because of the retrospective order  $p=1$ .

doi:10.1371/journal.pone.0115664.t003

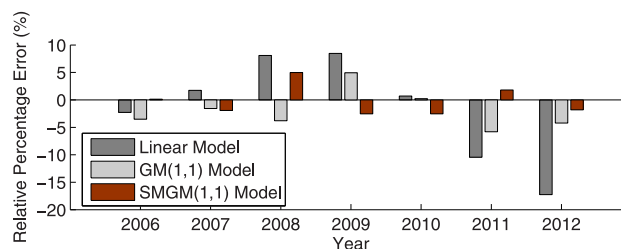


**Fig. 2. Comparison of incidence rates and simulated values among the three different prediction models for dysentery.**

doi:10.1371/journal.pone.0115664.g002

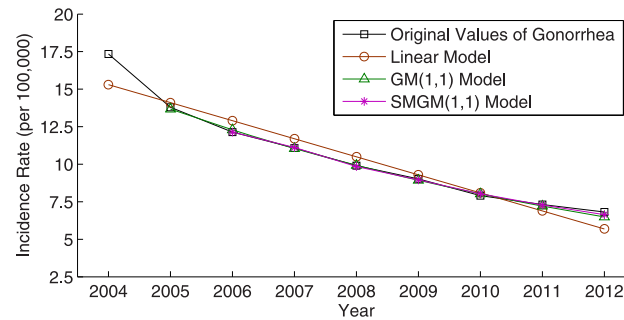
data. Through conversion of the original sequence, the model establishes a regression equation based on the regular generated sequence. Consequently, prediction of the dynamic development trend of diseases can be conducted using the regression equation. The GM(1,1) model is gradually developing into a common analytical method for both medicine and public health [22, 36, 37]. The grey self-memory coupling model is established on the basis of the conventional GM(1,1) model by combining it with the self-memory principle of dynamic systems. The coupling model can reflect the macroscopic development laws of disease systems based on the GM(1,1) model, and further explores their microscopic fluctuating laws with the help of the self-memory principle. The excellent prediction performance of the coupling model is supported by the systematic self-memorization equation containing multiple time-point initial fields instead of only single time-point initial fields. The equation overcomes the weakness of being sensitive to initial values of the conventional GM(1,1) model and takes full advantage of the system information contained in historical data [42].

In this paper, the conventional GM(1,1) model and the SMGM(1,1) model were utilized to predict the incidence rates of three representative infectious diseases in China. Analysis of simulation results indicated that the SMGM(1,1) model possesses more superior predictive performance than the conventional



**Fig. 3. Relative percentage error distribution of the three different prediction models for dysentery from 2006 to 2012.**

doi:10.1371/journal.pone.0115664.g003

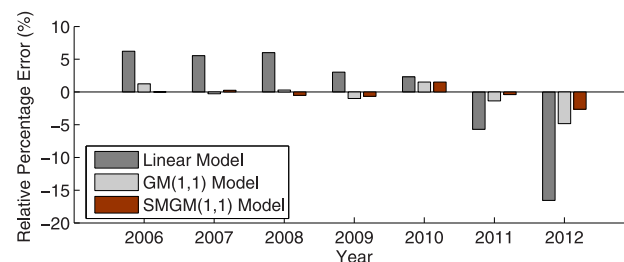


**Fig. 4. Comparison of incidence rates and simulated values among the three different prediction models for gonorrhoea.**

doi:10.1371/journal.pone.0115664.g004

GM(1,1) model. Therefore, the SMGM(1,1) model was selected to predict the future incidence rates of three infectious diseases in China. The incidence rate is the most straightforward and objective index, which evaluates the effect of prevention and control measures, because the numerator of the incidence rate is the new morbidity number within a time of resignation. The model significantly reduces prediction error of future incidence rates of infectious diseases. The combination of self-memory component and the grey system model can obtain encouraging prediction performances. This coupling model is a practical forecasting tool for infectious diseases with characteristics of stable epidemic factors, because its requirements of the original data are not as strict as those of other common statistical models. Therefore, the novel coupling prediction model can be used to predict the regularity of occurrence and development of infectious diseases and to identify their dynamic tendencies, thus providing a scientific basis for the prevention and control of infectious diseases.

Results of our simulations show that the novel grey self-memory coupling prediction model has improved the prediction accuracy of incidence rates, especially for short-term prediction. However, the model has certain limitations. First, when the grey self-memory coupling model is applied to long-term prediction, the prediction accuracy may decline. For this reason, the model needs to be further amended [49]. Over time, unknown disturbance factors may enter



**Fig. 5. Relative percentage error distribution of the three different prediction models for gonorrhoea from 2006 to 2012.**

doi:10.1371/journal.pone.0115664.g005

the system and exert unpredictable influences. Therefore, to forecast the long-term incidence rates of infectious diseases, the latest data should be utilized to amend the model to generate a new information model. Old data, which have minor effects on system development, may be discarded to improve the prediction accuracy. The spread of infectious diseases is inevitably influenced by natural, social, and environmental factors [50]. As a result, the most suitable model should be constructed by synthetically considering and systematically analyzing relevant influencing factors [22]. Prevention and early warning of infectious diseases can then be performed successfully using the appropriate prediction models.

## Acknowledgments

The authors wish to thank the editors and anonymous reviewers, whose constructive comments and suggestions were helpful in strengthening the presentation of this paper.

## Author Contributions

Conceived and designed the experiments: XG SL. Performed the experiments: XG LW. Analyzed the data: XG LW. Wrote the paper: XG SL LW LT.

## References

1. **Gonzalez-Castillo J, Candel FJ, Julian-Jimenez A** (2013) Antibiotics and timing in infectious disease in the emergency department. *Enfermedades Infecciosas y Microbiología Clínica* 31: 173–180.
2. **Thomas S, Luxon BA** (2013) Vaccines based on structure-based design provide protection against infectious diseases. *Expert Review of Vaccines* 12: 1301–1311.
3. **Brouqui P** (2009) Facing highly infectious diseases: new trends and current concepts. *Clinical Microbiology and Infection* 15: 700–705.
4. **Zhang L, Wilson DP** (2012) Trends in notifiable infectious diseases in China: implications for surveillance and population health policy. *PLoS One* 7: e31076.
5. **World Health Organization** (2013) World health report 2013.
6. **Apisarnthanarak A, Mundy LM** (2006) Infection control for emerging infectious diseases in developing countries and resource-limited settings. *Infection Control and Hospital Epidemiology* 27: 885–887.
7. **Andrus JK, Solorzano CC, de Oliveira L, Danovaro-Holliday MC, de Quadros CA** (2011) Strengthening surveillance: confronting infectious diseases in developing countries. *Vaccine* 29: D126–D130.
8. **Osei FB, Duker AA, Stein A** (2012) Bayesian structured additive regression modeling of epidemic data: application to cholera. *Bmc Medical Research Methodology* 12: 118.
9. **Nunes B, Natario I, Carvalho ML** (2011) Time series methods for obtaining excess mortality attributable to influenza epidemics. *Statistical Methods in Medical Research* 20: 331–345.
10. **Xu R** (2012) Global dynamics of an SEIS epidemic model with saturation incidence and latent period. *Applied Mathematics and Computation* 218: 7927–7938.
11. **Liu ZJ** (2013) Dynamics of positive solutions to SIR and SEIR epidemic models with saturated incidence rates. *Nonlinear Analysis: Real World Applications* 14: 1286–1299.



12. **Ma JQ, Wang LP, Qi XP, Shi XM, Yang GH** (2007) Conceptual model for automatic early warning information system of infectious diseases based on Internet reporting surveillance system. *Biomedical and Environmental Sciences* 20: 208–211.
13. **Dopson Stephanie A** (2009) Early warning infectious disease surveillance. *Biosecurity and bioterrorism: biodefense strategy, practice, and science* 7: 55–60.
14. **Hall IM, Gani R, Hughes HE, Leach S** (2007) Real-time epidemic forecasting for pandemic influenza. *Epidemiology and Infection* 135: 372–285.
15. **Hu MG, Li ZJ, Wang JF, Jia L, Liao YL, et al.** (2012) Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. *PloS One* 7: e38978.
16. **Getachew Y, Janssen P, Yewhalaw D, Speybroeck N, Duchateau L** (2013) Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models. *Statistics in Medicine* 32: 3224–3233.
17. **Bhatnagar Sunil, Lal Vivek, Gupta Shiv D, Gupta Om P** (2012) Forecasting incidence of dengue in Rajasthan, using time series analyses. *Indian Journal of Public Health* 56: 281–285.
18. **Zhang XY, Liu YY, Yang M, Zhang T, Young AA, et al.** (2013) Comparative study of four time series methods in forecasting typhoid fever incidence in China. *PloS One* 8: e63116.
19. **Nishiura H** (2011) Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009). *BioMedical Engineering OnLine* 10: 15.
20. **Nishiura H, Ejima K, Mizumoto K, Nakaoka S, Inaba H, et al.** (2014) Cost-effective length and timing of school closure during an influenza pandemic depend on the severity. *Theoretical Biology and Medical Modelling* 11: 5.
21. **Ren Y, Ding F, Suo S, Bu R-e, Zarlenga DS, et al.** (2012) Incidence rates and deaths of tuberculosis in HIV-negative patients in the United States and Germany as analyzed by new predictive model for infection. *PLoS One* 7: e42055.
22. **Shen XJ, Ou LM, Chen XJ, Zhang X, Tan XR** (2013) The Application of the Grey Disaster Model to Forecast Epidemic Peaks of Typhoid and Paratyphoid Fever in China. *PLoS One* 8: e60601.
23. **Hsin-Min L, Zeng D, Hsinchun C** (2010) Prospective infectious disease outbreak detection using Markov switching models. *IEEE Transactions on Knowledge and Data Engineering* 22: 565–577.
24. **Yaesoubi R, Cohen T** (2011) Generalized Markov models of infectious disease spread: a novel framework for developing dynamic health policies. *European Journal of Operational Research* 215: 679–687.
25. **Padhi R, Bhardhwaj JR** (2009) An adaptive drug delivery design using neural networks for effective treatment of infectious diseases: a simulation study. *Computer Methods and Programs in Biomedicine* 94: 207–222.
26. **Sree Hari Rao V, Naresh Kumar M** (2010) Estimation of the parameters of an infectious disease model using using neural networks. *Nonlinear Analysis: Real World Application* 11: 1810–1818.
27. **Guo H, Xiao XP, Forrest J** (2013) A research on a comprehensive adaptive grey prediction model CAGM(1,N). *Applied Mathematics and Computation* 225: 216–227.
28. **Cui J, Liu SF, Zeng B, Xie NM** (2013) A novel grey forecasting model and its optimization. *Applied Mathematical Modelling* 37: 4399–4406.
29. **Liu SF, Forrest J, Yang YJ** (2013) Advances in Grey Systems Research. *The Journal of Grey System* 25: 1–18.
30. **Xie NM, Liu SF, Yang YJ, Yuan CQ** (2013) On novel grey forecasting model based on non-homogeneous index sequence. *Applied Mathematical Modelling* 37: 5059–5068.
31. **Wu LF, Liu SF, Yao LG, Yan SL** (2013) The effect of sample size on the grey system model. *Applied Mathematical Modelling* 37: 6577–6583.
32. **Li J, Wang BY, Zhang BS** (2006) Application of improved grey prediction model to petroleum cost forecasting. *Petroleum Science* 3: 89–92.
33. **Li GD, Wang CH, Masuda S, Nagai M** (2011) A research on short term load forecasting problem applying improved grey dynamic model. *International Journal of Electrical Power & Energy Systems* 33: 809–816.

34. **Li DC, Chang CJ, Chen CC, Chen WC** (2012) Forecasting short-term electricity consumption using the adaptive grey-based approach—an Asian case. *Omega* 40: 767–773.
35. **Hsu LC** (2003) Applying the grey prediction model to the global integrated circuit industry. *Technological Forecasting and Social Change* 70: 563–574.
36. **Lin WZ, Fang JA, Xiao X, Chou KC** (2012) Predicting secretory proteins of Malaria Parasite by incorporating sequence evolution information into pseudo amino acid composition via grey system model. *PLoS One* 7: e49040.
37. **Ding F, Zarlenga DS, Qin C, Ren X** (2011) A novel algorithm to define infection tendencies in H1N1 cases in Mainland China. *Infection, Genetics and Evolution* 11: 222–226.
38. **Cao HX** (1993) Self-memorization equation in atmospheric motion. *Science in China (Series B)* 36: 845–855.
39. **Song B, Lee C H** (2010) A Favré averaged transition prediction model for hypersonic flows. *Science China Technological Sciences* 53: 2049–2056.
40. **Liu XP, Xie XB, Luo YZ** (2010) Self-memorization model of dynamic system for predicting nonlinear displacement of slopes. *Chinese Journal of Geotechnical Engineering* 32: 1535–1542. (in Chinese)
41. **Wang W, Su JY, Hou BW, Tian J, Ma DH** (2012) Dynamic prediction of building subsidence deformation with data-based mechanistic self-memory model. *Chinese Science Bulletin* 57: 2171–2176.
42. **Fan XH, Zhang Y** (2003) A novel self-memory grey model. *Systems Engineering - Theory and Practice* 23: 114–117. (in Chinese)
43. **Chen XD, Xia J, Xu Q** (2009) Differential hydrological grey model (DHGM) with self-memory function and its application to flood forecasting. *Science in China Series E: Technological Sciences* 39: 341–350. (in Chinese)
44. **Guo XJ, Liu SF, Fang ZG** (2014) Self-memory prediction model of interval grey number based on grey degree of compound grey number. *Systems Engineering and Electronics* 36: 1124–1129. (in Chinese)
45. **Gao L, Zhang Y, Ding GY, Liu QY, Zhou MG, et al.** (2014) Meteorological variables and bacillary dysentery cases in Changsha city, China. *The American Society of Tropical Medicine and Hygiene* 90: 697–704.
46. **Jakopanec I, Borgen K, Aavitsland P** (2009) The epidemiology of gonorrhoea in Norway, 1993–2007: past victories, future challenges. *BMC Infectious Diseases* 9: 33.
47. **Ministry of Health of the People's Republic of China** (2013) China health statistical yearbook of 2013. Beijing: Peking Union Medical College Press. (in Chinese)
48. **Cao HX** (2002) Self-memory principle of dynamic system: forecast and computing applications. Beijing: Geological Publishing House. (in Chinese)
49. **Ding F, Zarlenga DS, Ren Y, Li G, Luan J, et al.** (2011) Use of the D-R model to define trends in the emergence of Cefotaxime-resistant *Escherichia coli* in China. *PLoS One* 6: e27295.
50. **Watkins RE, Eagleson S, Hall RG, Dailey L, Plant AJ** (2006) Approaches to the evaluation of outbreak detection methods. *BMC Public Health* 6: 263.