Proceedings of Business and Economic Studies

Research Article

Big Data Interprets US Opioid Crisis

Zidong Wang[#], Poning Fan[#]

Donghua University, Shang Hai 201620, China

[#]These authors contributed equally to this work

Abstract: Since 2010, there has been a new round of drug crises in the United States. The abuse of opioids has led to a sharp increase in the number of people involved in drug crimes in the United States. There is an urgent need to explore solutions to the drug crisis in the United States. In this paper, the model of in-depth analysis is established under the condition of obtaining the opioid data and the influence factor data of the large sample of five state[1]. In the first part, we use the Highway Safety Research Institute model based on the differential equation model to predict the initial value, find the initial position of the drug transfer, and obtain the curve of the number of different groups over time by fitting the data, so that the curves can be predicted the changing trends of the groups in the future. It was found that in Kentucky State, the county's most likely to start using opioids were Pike and Bale. In Ohio, the county's most likely to start using opioids are Jackson and Scioto. In Pennsylvania State, Mercer and Lackawanna are the counties most likely to start using opioids. Martinsville and Galax are the counties where Virginia State is most likely to start using opioids. Logan and Mingo are the counties where West Virginia State is most likely to start using opioids. In the second part, the gray prediction model is used to further analyze the time series of each factor, the maximum likelihood estimation method is used to obtain the weight of each factor, and the weight coefficient matrix is used to simulate the multivariate regression equation, and the factors that have the greatest influence on opioid abuse are educational background and family composition. In the third part, the hypothesis test model of two groups (the data type is proportional) is used to verify the difference between the influence factors (including the predicted values) in the first two parts of the states, thus verifying the feasibility between them. At the same time, we put forward a few suggestions to combine the current situation in the United States with the CDC data. We believe that in order to address the opium crisis, the U.S. government needs to strengthen not only oversight of doctors' prescriptions, but also make joint efforts of all sectors of society to fundamentally reduce the barriers to the use of opioids.

Keywords: *Highway Safety Research Institute model; synthetic drug, data itting; gray prediction; hypothesis test; antidrug advice*

Publication date: December, 2020

Publication online: 30 December, 2020

Corresponding Author: Zidong Wang, 13598090952@139.com

0 Introduction

The purpose of this paper is to make a mathematical analysis of the opioid abuse in the United States by combining the data and infer the possible causes of the opioid crisis. In view of these reasons and the current national conditions of the United States, some measures and opinions are proposed. All the inferences in this paper are based on the statistical analysis of the data, so the conclusions may be more authentic. The analysis of drug abuse in the United States also combines some social factors in the United States. This paper belongs to the big data analysis class.

1 Problem 1

1.1 Highway Safety Research Institute (HSRI) model

Based on the existing virus propagation model, thought propagation and flow model, the HSRI model is

established to study the cyclical problem of drug flow between states [Figure 1 and Table 1].

Assume that, all parameters are non-negative constants. The basic regeneration number of this model (HSRI model) is:

$$R_0 = \frac{\beta\Lambda}{\gamma+\delta_1} + \frac{\gamma\varepsilon}{(\gamma+\delta_1)(\varepsilon+\delta_2)}$$

During the steady-state phase of propagation, the relative number of these numbers reaches a stable distribution ratio so that the following differential equations can be used for characterization:

$$\frac{dH}{dt} = \alpha(1-H-S-R)H - (\beta+\varepsilon)H$$

$$\frac{dS}{dt} = \beta H - \gamma S$$

$$\frac{dR}{dt} = \gamma S + \varepsilon H - \eta R$$

The initial value set for the function S,h,r for time (in years here) is:

$$D = \{(H,S,R) \mid H,S,R \geq 0 \,\&\, h+s+r \leq 1\}$$

When and only when $I = 1$, $r + h + s = 1$

By solving this differential equation group, h (t), S (t), and R (t) are obtained, to seek the uniform unity of the form of solution, so the constant terms are omitted in the lower formula, and then, the accuracy and reliability of the prediction can be increased according to the actual number of state data in the process of solving the parameters. The following are the expression of each function:

$$H = e^{\frac{\alpha k}{1-\alpha}t} \qquad (*)$$

$$S = \frac{\beta}{\gamma}e^{\frac{-\alpha k}{1-\alpha}t} - \frac{1}{\gamma}e^{-kt}$$

$$R = (\gamma\beta+\varepsilon)e^{\frac{-\alpha k}{1-\alpha}t} - \frac{1}{\gamma}e^{-kt} - e^{-\eta t}$$

$$\Theta = \frac{\alpha}{\beta+\varepsilon}$$

If the number of drug communicators does not exceed the threshold, the control of opioids is controlled by the government.

However, when the number of s, h exceeds the alert limit (0.05), the trade and spread of heroin will move in an uncontrollable direction, with a surge in numbers, so monitoring through big data is something the U.S. government must do.

Find the values for each parameter as follows:

$$\alpha = 0.1, \beta = \varepsilon = 1, \gamma = 0.618, k = 1.9375$$

The initial value (2010) of the percentage of the total number of people in each state is approximated by the data given in 2010, Figure 2:

$$s(0) = 0.68\%, r(0) = 3.3\%, h(0) = 1.13\%\partial$$

Table 1. Symbol description explanation in the model

I	Ignorant person, good citizens who do not have any propensity to use drugs
H	Lurker, a population with a high risk of drug addiction.
S	Communicators, who are already addicted to drugs and can infect some of them to some extent
R	The removal of the person, the people who detox between the Lurker and the communicator
ε	The probability that heroin users should stop drug treatment and relapse, that is, the exchange ratio between the emigration and the Lurker
γ	Indicates the rate at which heroin users enter drug treatment to become better
s	Communicators account for the proportion of the state's total population
h	The proportion of the total population of the state that the Lurker accounts for
r	The proportion of the total population of the state as a migrant

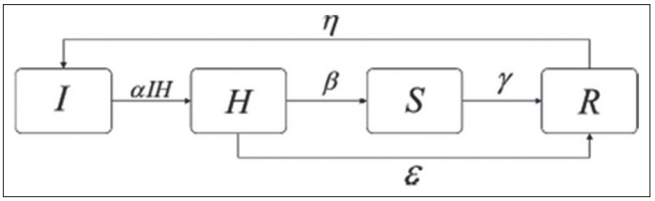

Figure 1. Highway Safety Research Institute model diagram

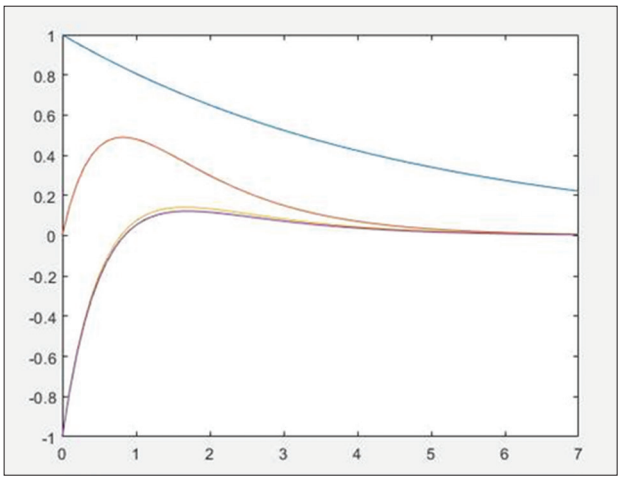

Figure 2. *S*, *t*, and *r* fitting curve of the change overtime

The number of communicators grew rapidly and peaked a year later, but the peak was not sustained, gradually fell to a lower level over the next 5 years.

1.2 Prediction based on Gray GM (1,1) model

The contribution of each factor to drug use (opioids or heroin) may be significantly fluctuating by extreme data or the impact of a particular environment, so the gray projection method is used to add up the original data series[2], thus maximizing the problem caused by less historical data and insufficient reliability. Here, the number of influencing factors n takes 5.

Given observation data:

$$\mathrm{x}^{(0)} = \left\{ x^{(0)}(1) \quad \mathrm{x}^{(0)}(2) \quad ... \quad x^{(0)}(n) \right\}$$

To weaken the randomness of the original time series, the original time series needs to be processed before the gray prediction model is established, and the time series after data processing is called the generating column. Moreover, a cumulative method is used to obtain the generation sequence.

After a cumulative: $x^{(1)} = \left\{ x^{(1)}(1) \quad x^{(1)}(2) \quad ... \quad x^{(1)}(n) \right\}$

Set $\mathrm{X}^{(1)}$ to meet the first-order ordinary differential equations

$$\frac{d\mathrm{x}^{(1)}}{dt} + ax^{(1)} = u$$

Where a is a constant, called the development of gray number, called endogenous control ash number, is a specific constant input to the system. This equation satisfies the initial conditions,

When $t = t_{0,}\ x^{(1)} = x^{(1)}(t_0)$ the formula is:

$$x^{(1)}(t) = [x^{(1)}(t_0 - \frac{u}{a}), e^{-\alpha(t-\tau_0)} + \frac{u}{a}$$

$X(0)$ is set to the academic qualifications, the family structure ratio, the mother tongue type, the United States citizen share these four factors, can get the forecast value. $x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\}(n = 7)$

$\forall i, x_i \geq 0 \,\&\, \sum_{i=1}^{n} x_i = 1$ and meet: $x = \{x_i \mid i = 1,2,3,\cdots,n\}$, the gray entropy function Xi called $\Omega(x)$ as the sequence value x is the information attribute, when $xi = 0$, $\Omega(x) = 0$, when X_i tends to change evenly isometric, the ash entropy increases. When completely uniform, that is, when the contribution of each factor to the rate of opioid consumption is equal, the gray entropy is taken to the maximum value, which is:

$$\Omega(x)_{\max} = \ln n$$

A called resolution, 0<call. The smaller the Rho, the greater the difference between the related numbers, the stronger the differentiation ability, here to take ρ = 0.5.

$$\Delta^{(0)}(i) = | x^{(0)}(i) - \hat{x}^{(0)}(i) |$$

$$\Gamma(i) = \frac{\Delta^{(0)}(i)}{x^{(0)}(i)} \times 100\%$$

The original influence factor sequence, relative error sequence, and absolute error sequence are calculated, respectively, according to the above two formulas.

Discrete values for peer interval sampling (Note when $t_0 = 1$) are $x^{(1)}(k+1) = [x^{(1)}(k+1) = [x^{(1)}(1) - \frac{u}{a}]e^{-ak} + \frac{u}{a}$ the way to model Gray is to accumulate sequences once, the constant *a* and *u* are estimated by the least square's method. Use $x^{(1)}$ (1) as the initial retention value, and the rest is replaced by difference to obtain the following formula:

Replace the $\mathrm{X}^{(i)}$ (i) with the $\frac{1}{2}[x^{(i)}(i) + x^{(i)}(i-1)], (i = 2,3,...,N)$. Overwrite matrix expressions are available:

$$\text{Make } y = (x^{(0)}(2), x^{(0)}(3), \cdots, x^{(0)}(N))^{\mathrm{T}}.$$

The matrix form is: $Y = BU$, estimation with least squares method $\hat{U} = \begin{bmatrix} \hat{a} \\ \hat{u} \end{bmatrix} = (B^T B)^{-1} B^T y$ substitute estimates â and û, available response equations: $\hat{x}^{(1)}(k+1) = \left[x^{(1)}(1) - \frac{\hat{u}}{\hat{a}} \right] e^{-\hat{a}k} + \frac{\hat{u}}{\hat{a}}$, this allows you to calculate the corresponding data.

To obtain the predicted value of the original sequence, the predicted value of the generated series needs to be reduced to the original value.

$$\hat{X}^{(0)}(k) = \hat{X}^{(1)}(k) - \hat{X}^{(1)}(k-1)$$

This is an exponential growth model, and when making predictions, the forecast results for the past year should be accurate, but the forecast error for subsequent years will gradually increase[3], to improve the broad applicability of the predictive model, we have made the following improvements: Using the method of multivariate regression analysis, the relative weight of each factor is obtained, and the weight matrix is taken into the predictive function as the correction coefficient, which makes accurate prediction, and avoids the damage of the operation to the information integrity of each state.

x = arrange (−1,1,0.02) y = ((x*x−1)**2+2)*(sin(x*3)+0.7*cos(x*1.2)) the fitting function given in the figure above represents four influencing factors (education, family, and language, whether American citizen) [Figure 3].

Based on the calculation of the gray model, we predicted the total amount of opioid drug reports in the states of 2018, 2019, and 2020. In this regard, we suggest that the United States government should take action as soon as possible to curb the proliferation of opioid use. In terms of data, the trend in total opioid reporting in five states varies, with KY, OH shows an upward trend [Table 2], which is a very dangerous signal. The U.S. government has introduced policies in recent years to curb the proliferation of opioids, but if the amount of opioid reporting does not decrease[4,5]. In this case, the U.S. government needs to introduce other effective policies as soon as possible to avoid the situation becoming more serious.

In VA State, although the 3-year forecast does not reach the threshold of 35,000, the data are very close and the government also needs to be aware of it. Among them, the most important attention is Norton, Martinsville, and Galax three counties, it is recommended that the United States government to establish a drug information sharing network system under the premise of the three counties to carry out high-frequency spot checks to ensure that the annual value is at the threshold.

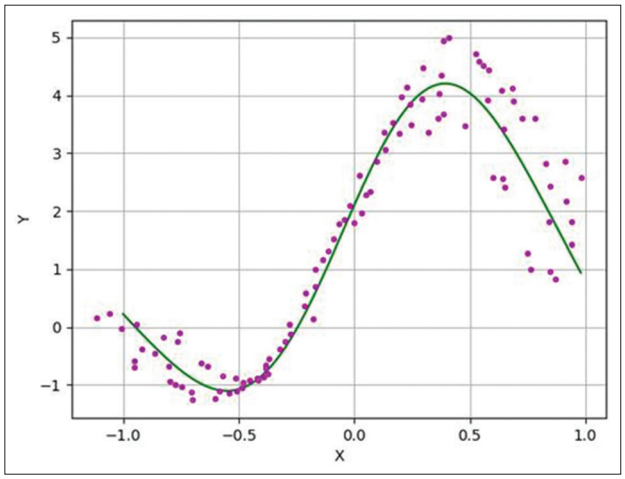

Figure 3. Least squares fitting graph

2 Problem 2

The flow propagation process of drugs between state and state is analyzed using the fitting prediction model of function.

2.1 Using multivariate factors to analyze the weight of legal influencing factors

On the basis of the existing model and discussion of the first question, the method of factor analysis is introduced to make the weighted prediction so that the model is further close to the actual situation.

$$Y = \beta_0 + \beta_1 x + \beta_2 x + \beta_3 x + \beta_4 x + \mu$$

β_i is the partial regression coefficient, (i = 0, 1, 2, 3, 4) indicates the rate of opioid consumption, respectively, the number of educated (undergraduate and above) is the percentage of the state, number of family integrity as a percentage of total, the proportion of people whose mother tongue is English is the total number of people in the state, and U.S. citizens account for the total number of people in the state. The following matrices are various influencing factors (year 2012–2016) load matrix of weights, the principal component analysis method and the maximum likelihood estimation method are given:

$$\Lambda = [\sqrt{\lambda_1}\eta_1, \sqrt{\lambda_2}\eta_2, \cdots, \sqrt{\lambda_m}\eta_m] \quad (1)$$

Which λ is the eigenvalue of the correlation coefficient matrix R,

$$Q = \sum_{i=1}^{4} (y_i - \beta_0 - \beta_i x_i)^2$$

For each factor that affects the number of opioid drug dosage, the partial derivative is calculated, and the stable point with partial derivative is 0, wherein the sample observation point (x_i, y_i) is selected, and several factors selected here are several kinds of factors which are different from the influence mode obtained by cluster analysis, so the sample is not only strong enough to be representative but can also truly reflect the situation of opioid consumption in these states of the United States.

Table 2. OH state in the top three counties with opioid smoking rates in 4 years

2014		2015		2016		2017	
Countiy	**Rate**	**Countiy**	**Rate**	**Countiy**	**Rate**	**Countiy**	**Rate**
Jackson, OH	175.3	Jackson, OH	151.8	Jackson, OH	133.8	Jackson	112.5
Washington, OH	147.5	Washington, OH	133.7	Washington, OH	119.9	Washington	102.7
Jefferson, OH	138.7	Jefferson, OH	121.8	Jefferson, OH	111.9	Pike	93.6

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{4}(y_i - \beta_0 - \beta_i x_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{4}(y_i - \beta_0 - \beta_i x_i)x_i = 0$$

The slope and intercept of single factor variables are obtained by the above equations.

$$\beta_i = \frac{\sum_{i=1}^{4}(y_i - \overline{Y})(x_i - \overline{X})}{\sum_{i=1}^{4}(X_i^2 - n\overline{X}^2)}$$

$$\beta_0 = \overline{Y} - \beta_i \overline{X}$$

The fitting curve is obtained to minimize the square sum of the residual of the actually measured data y_i and the corresponding estimate on the fitting line.

$$\eta_j = \hat{\beta}_j \frac{\overline{X}_j}{\overline{Y}} \quad (2)$$

In formula, η_j indicates that the relative importance order of the variables is calculated by the range of the respective average of the different influencing factors, each change of the independent variable 1%, and the extent of the change in the percentage of opioid smoking is: Education (University and above) > Family composition (family is complete) > native language type > is a U.S. citizen.

The forecast values for the next 3 years are shown in Table 3:

Bring the obtained average to the type $E(\overline{\beta}) = \hat{\beta}$ validation, in cases where the error does not exceed 0.05, it is considered that the predicted value of each factor statistic is the unbiased estimate of the overall sample.

2.2 The distribution of the incidence of drug use simulated by MATLAB simulation [Figure 4]

As we can see from the figure, the curve on the graph represents the contour line, and the deeper the cold hue represents the more controllable the use of opioids in the region. Conversely, the deeper the warm hue, the worse the abuse of opioids in the region. We can find that, combined with 10 years of data, Ohio is much worse than other places, and its color basically floats between orange and light yellow; the color scale of the West Virginia State in 100–200 of the blue to the dark blue range, so the US government has to focus its efforts on Ohio rather than the West Virginia State.

Table 3. Forecast values for the next 3 years

Year	KY	OH	PA	VA	WA
2018	27,070	114,592	72,159	32,784	4807
2019	27,490	116,405	71,095	34,439	4628
2020	27,810	116,782	70,668	34,739	4369

2.3 Analysis and recommendations

Prescription drug abuse has become one of the most severe and fastest-growing drug problems in the United States in recent years. According to the National Drug Use and Health Survey of the United States Drug Abuse and Mental Health Services Authority, in 2016, >11 million people in the United States abused opioid prescription drugs, nearly 1 million people used heroin and 2.1 million had opioid use disorders. Therefore, finding the best balance between the control and rational use of opioids becomes an urgent problem. Administrators, law enforcement officials, and drug addiction groups often attribute addiction to the blind expansion of the market by drug companies and the inaction of drug regulators. However, in-depth analysis of different types of drug abuse and the degree of abuse of people, the reasons for abuse are also different. The vigorous promotion of pharmaceutical enterprises: The U.S. pharmaceutical industry and economic woes have largely contributed to an increase in the number of drug overdose deaths across the United States, and the annual increase in the investment budget for drug advertisements has attracted a growing number of patients or subhealth groups who do not fully understand their condition and side effects of drugs, a group in this model for the Lurker to be infected (H), drug companies, on the other hand, are like communicators in models (S).

2.3.1 Processing scenarios

Strengthening the requirement for drug companies to provide post-market data on the effects of long-term use

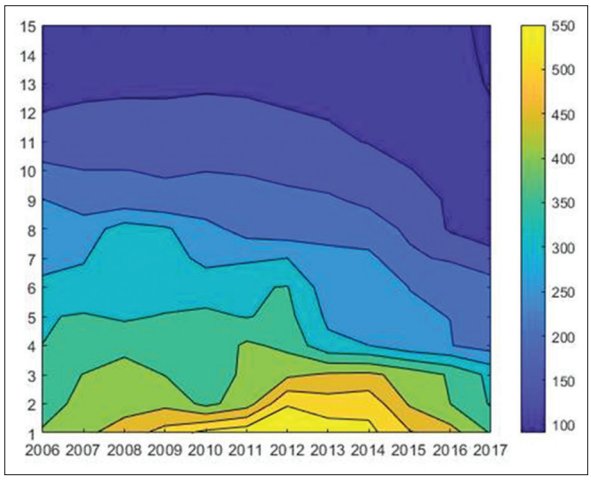

Figure 4. Probability density distribution of opioid users

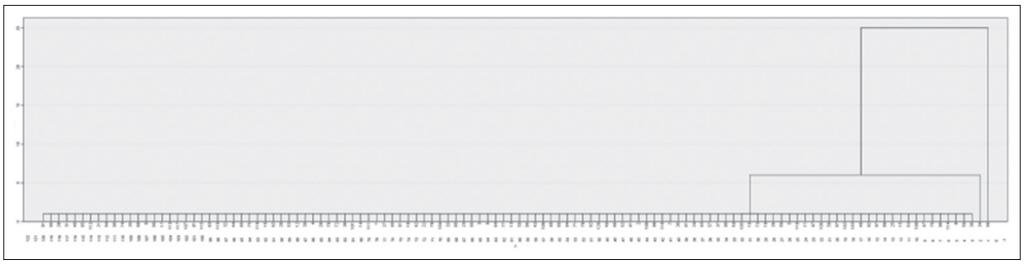

Figure 5. Use a tree diagram of an average join (intergroup)

Table 4. OH summary of state influencing factors

Year	Qualifications	Family composition	Language	Whether US citizens	Opiates drugs report total
10.00	17.44	69.25	95.64	50.17	70,999.00
11.00	17.68	69.15	95.54	48.72	71,282.00
12.00	17.94	68.87	95.47	48.44	85,415.00
13.00	18.23	68.67	95.42	49.58	93,747.00
14.00	18.53	68.35	95.42	50.11	101,423.00
15.00	18.85	67.94	95.39	50.63	109,150.00
Average value	19.26	67.72	95.41	51.39	115,276.00
Z	−18.25	2.72	14.36	−33.08	

of er/la opioids provides better evidence of the serious risks of misuse, abuse, and long-term use of opioids, which can be networked to track Lurker (H) and communicator (S) and establish information-sharing mechanisms with pharmacies, this group of people to buy opioids more stringent audit.

The positive effects of opioids are equally high, and their main effect is to be able to treat moderate-to-severe pain, such as pain caused by cancer. The analgesic effect of opioid efficacy is good: If you take a low dose, you can achieve the desired analgesic effect. However, the side effect is the strong addictive nature. Patients with excessive medication may turn to heroin later in the latter stages so that heroin users in some states increase year after 2010, while opioid users experience negative growth, which is a possible cause. In addition, in the socioeconomic data, indicators given by the United States Bureau of Statistics, residents' mother tongue, academic background, family membership structure, family per capita income, and other factors are also strongly correlated with the drug infection. The method of tree clustering analysis using SPSS can also sift out these important influencing factors, as shown in Figure 5.

3 Problem 3

Taking Kentucky State and Ohio as examples [Table 4], to verify the correctness of the data obtained from the model, two states are selected as two overalls, comparing the effects of various influencing factors on opioid consumption rate in two states, the original hypothesis H_0 is neutral, and it is considered that these effects are not different between the states. H_1 as an alternative hypothesis considers that the effect of the factors on opioid rates varies within each state and that this bias may be derived from whether the state's legal policies can deter the drug, or the local family structure and geographical environment can also have an impact[6]. The following method of hypothesis test is used to verify the accuracy of the model established by the first two questions:

The following method of hypothesis test is used to verify the accuracy of the model established by the first two questions:

$$H_0 : \pi 1 - \pi 2 = 0$$

$$H_1 : \pi 1 = \pi 2$$

Test of the difference of two overall proportions:

Because the influencing factors enumerated in the model are subject to two distributions (that is, by the measure of "true" or "false"), the two overall are Kentucky State and Ohio, respectively. The proportion of units with certain characteristics in the sample is π_1, π_2, respectively, but they are unknown quantities and can be replaced by sample proportional p_1, p_2.

Under the condition that the original hypothesis is established, the best variance is P (1-P), where p is the proportional estimate obtained by merging two samples. That

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n1} + \frac{1}{n2})}}$$

X_1 represents the number of units in a sample N_1 that has certain characteristics and X_2 represents the number of units in the sample N_1 that has certain characteristics. Under large samples, the expression of statistic Z is:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n1} + \frac{1}{n2})}}$$

4 Conclusion

A lot of people worry that big data on health and substance abuse will involve all aspects of life and are prone to infringement of citizens' privacy, so appropriate legal safeguards should be put in place before large data can be used on a large scale. It is conceivable that, once norms and laws are in place, the collection and use of big data on drug abuse will advance by leaps and bounds, completely breaking the limits of traditional data and promoting the rapid development of drug abuse research and practice.

References

[1] Center for Disease Control and Prevention [EB/OL]. Available from: https://www.cdc.gov/drugoverdose/. [Last accessed on 2019 Jan 26].

[2] Qiang SG, Yue L. New application of grey prediction in border drug intelligence analysis. Inf J 2011;30 S 1:21-22.

[3] Fen YD, Long YQ, Chen LW. Calculation and implementation of hazardous chemical leakage area based on Gaussian diffusion model. Comput Appl Chem 2012;29:195-199.

[4] Hui YY, Li XX, Zhu Z. Overview of opioid abuse in the United States and its control measures. China Drug Alert 2017;14:746-751.

[5] Li S. Overview of opioid abuse in the United States and its control measures. China J Drug Abuse Prev Treat 2018; 24:219-224.

[6] De L. Big data construction in the drug abuse crisis in the United States. Peoples Rule Law 2018;Z1:157-158.