REPORTS OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
ISSN 2224-5227
Volume 5, Number 315 (2017), 5 – 11

B. Zhussupov[a], S. Hermosilla[b], A. Terlikbayeva[a], A. Aifah[b], Z. Zhumadilov[c], T. Abildayev[d], T.Muminov[e], R.Issayeva[c]

[a]Columbia University Global Health Research Center of Central Asia; 102 Luganskogo Street, Almaty 050051, Kazakhstan;
[b]Columbia University in the City of New York; 722 West 168th Street, Room 507, Box 13, New York, NY 10032, United States of America;
[c]Center for Life Sciences Nazarbayev University; 5 KabanbayBatyr Street, Astana 010000, Kazakhstan;
[d]National Center for Tuberculosis in Kazakhstan; 5 Bekhodjin Street, Almaty 050059, Kazakhstan;
[e]Kazakhstan Association of TB Specialists;
Corresponding Author: BaurzhanZhussupov, 102 Luganskogo Street, Almaty 050051, Kazakhstan; baurzhan.zhussupov@gmail.com; tel: +7(727) 2646930; fax: ext. 112

TIME-SERIES ANALYSIS ON NEW TB CASES IN KAZAKHSTAN

Author contributions:
Zhussupov had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: Zhussupov, Hermosilla, Muminov, Aifah,Terlikbayeva. Acquisition of data, critical review of intellectual content and final approval of the version to be published: Zhumadilov, Abildayev, Issayeva. Analysis and interpretation of data: Zhussupov, Hermosilla, Aifah, Tinglin. Drafting of the manuscript: Zhussupov, Hermosilla, Aifah. Terlikbayeva. Statistical analysis: Zhussupov, Hermosilla. Obtained funding: Terlikbayeva, Zhumadilov, Abildayev, Issayeva, Muminov.

Abstract. Objectives. To evaluate apredictive national time series model for tuberculosis (TB) incidence constructed as the sum of the regional models in comparison with a model based only on national data.
Key words: Tuberculosis, ARIMA, incidence forecasting.

Study Design
We conducted a comparison of TB forecasting models based on Kazakh national health surveillance data from 2007 to 2013.
Methods
The autoregressive integrated moving average (ARIMA) models were constructed with the data on the monthly number of newly reported TB cases from 2007 to 2012 for all administrative regions and at the national level. The first national model was built based on national data only. The second national model was the sum of the regional models. Data from 2013 were used to test the performance of the models.
Results
Seasonal ARIMA$(0,1,1)(1,0,1)_{12}$ model demonstrated best fit to the national TB notification rates. Regional ARIMA models varied from region to region.Mean absolute percentage error (MAPE) of the sum of regional models was 7.3, whereas the model based only on national data had MAPE equaling 10.7.
Conclusions
The sum of the regional models is more accurate than the model based on national data only. The national ARIMA model and the regional models correctly reflect Kazakhstan's downward trend in notification of new TB cases. Improving forecasting models at a national level will help stem the multidrug-resistant tuberculosis epidemic.

Introduction

In 2012, there were 8.6 million new casesof tuberculosis (TB) and 1.3 million deaths from the disease registered worldwide, continuing a decade-long decline in the global trajectory of the disease.[1] Although

the TB incidence rate in Kazakhstan is also experiencing a downward trend, now at 73.5 per 100,000 in 2013 compared to 165 per 100,000 in 2002, Kazakhstan continues to havehighincidence rate of active TB and has one of the world's highest proportions of multi-drug resistance TB (MDR-TB) among new and previously treated TB cases. The estimated proportion of new TB-cases that that were MDR-TB was sixth highest in 2012 globally and Kazakhstan is considered as a high MDR-TB burden country.[1,2]

The registration of new TB cases provides a good estimate of the actual TB incidence rate,[3] which varies by season in many countries.[4-7] The timely prognosis of the disease is crucial in terms of planning and evaluating the subsequent public health response.

One of the well-established methods of predicting incidence, including seasonal incidence, is the Autoregressive Integrated Moving Average (ARIMA) model.[8,9] ARIMA uses national data to identify the optimal model for TB incidence, and then estimates its parameters to predict future cases of the disease. In addition to direct national estimation, another way to build a national model is through the summing of the constructed regional models. After the model is created, the final step is to assess its performance by comparing it against real data.[8,9]

ARIMA models create point estimates and error terms. The errors of an ARIMA model must be stationary and normally distributed:

$$Y(t) = \hat{Y}(t) + \varepsilon(t),$$

whereY(t) is the actual value of the series at time t; $\hat{Y}(t)$ is an estimate for the time t, and ε(t) is a normally distributed stationary error. Therefore, after summing independent ARIMA models, we also have a model in the same form.

Kazakhstan is an ideal location to test this hypothesis because the current TB system is consistent and doesn't fluctuate by region. The Kazakh National TB Center coordinates all TB activities including TB surveillance and TB case notification at the national level. Regional TB dispensaries work under the supervision of regional health departments and the National TB Center. TB treatment is free for all patients with legal residence, provided by local TB dispensaries, and funded by local or central government budgets. All TB cases, new and relapsed, are continually recordedin the National TB register.

Kazakhstan has 16 administrative regions, which consist of 14 oblasts and two cities-Almaty and Astana. In geographically large countries as Kazakhstan, regions differ in terms of geography, climate, population density, and level of economic development. Taking into account such heterogeneity we expect that regional ARIMA models predicting TB incidence rates may also vary from region to region, possibly leading to inconsistencies with the national optimal ARIMA model.

We hypothesize that the national model predicting TB incidence constructed as the sum of the regional models shows a better fit than the model that is based only on national data. To examine this hypothesis, we constructed three forecasting models after analyzing the seasonality of TB in Kazakhstan. The first forecasting model is a simple application of national monthly incidence of the previous year. This approach is intuitive and widely used as a prediction method in public health when other more sophisticated forecasting methods are not available. The second model is a model of the seasonal ARIMA model based on national data, and the third model is anaggregated national model comprised of multiple regional ARIMA models. To construct these models, we used new cases of TB occurring on a monthly basis from 2007 to 2012 and then tested their performance on National TB registry data from 2013.

Materials and Methods

Study design

We evaluated TB forecasting models based on Kazakh national health surveillance data from 2007 to 2013. Data on the monthly number of newly reported TB cases at the regional and national level were obtained from the National Center for Tuberculosis. To construct the ARIMA model, we used the data over six years from 2007 to 2012. Data from 2013 were used to test the performance of the models.

Determination of seasonality

To determine the seasonality of TB notification rates, we identified whether monthly rates differ from each other. To estimate the difference we used the analysis of variance of 12 groups of incidence rates, made up by the twelve months in a calendar year.

Model identification

A seasonal ARIMA model can be represented by ARIMA $(p, d, q)(p_{12}, d_{12}, q_{12})$. The main purpose of model identification is to determine p, d, q, p_{12}, d_{12}, q_{12} values, where p and p_{12} define the number of autoregressive and seasonal autoregressive terms; d and d_{12} are number of required differences, nonseasonal and seasonal; q and q_{12}, thenumber of nonseasonal and seasonal moving average parameters. The optimal model should produce estimations which provide the best fit to the actual data with the lowest number of parameters and allows stationarizing a time series. Selection of optimal ARIMA models was based on the minimization of Akaike's information criterion with a correction for finite sample sizes (AICc). Models were estimated in R using the package *forecast*.[10]

Diagnostic testing for residuals

Wechecked the adequacy of fitted models to the data by auto-correlation functions (autocorrelation and partial autocorrelation functions) of residuals to ensure the presence of small correlations. Histograms and Q-Q plots of errors were constructed to be sure that errors have normal distribution.

Forecasting

We conducted forecasting using the prediction equation of a known ARIMA model for 12 months of 2013. The 'forecast' function from the R package *forecast* uses existing time series data in an ARIMA model equation to forecast results. Performance of forecasting was evaluated by three measures: mean absolute deviation (MAD), root-mean-square error (RMSE) and mean absolute percentage error (MAPE).[8]

Results

Seasonality

Figure 1 presents data on newly registered TB cases by month from January 2007 to December 2012. The data demonstrate monthly variation. Figure 2 compares the average values by month for 6 years, the peak of case registration is in April (1778 cases) with minimums in September (1091 cases) and December (986 cases). Analysis of variance results indicate significant differences in the incidence of monthly TB cases: $F_{11,60} = 5.752$ ($p < 0.001$), which necessitates the use of seasonal (monthly) ARIMA model - $SARIMA_{12}$.

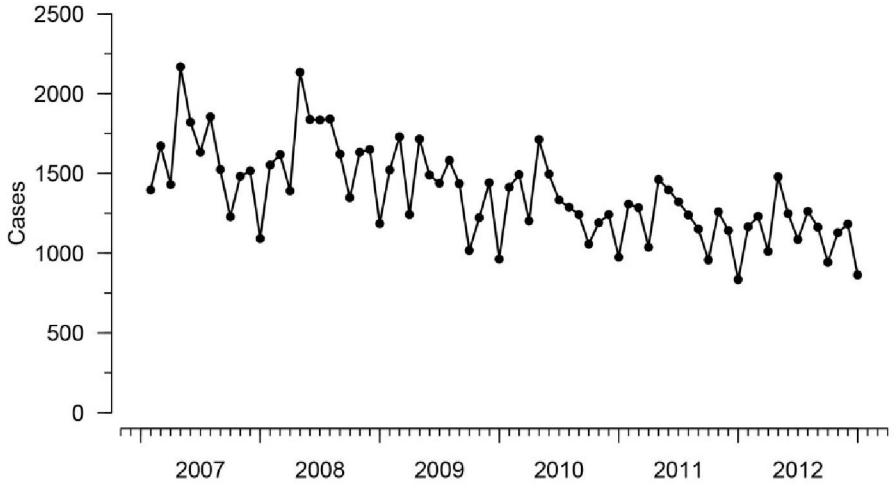

Figure 1 - Number of new TB registered cases in Kazakhstan, per month (2007-2012) (provided in the separate file)

Models

The first model, repetition of monthly incidence from previous year, can be recorded as $SARIMA(0,0,0)(0,1,0)_{12}$ model. The second model constructed based on the national TB notification data is a $SARIMA(0,1,1)(1,0,1)_{12}$ model. To build the third model, summing of regional models, we identified optimal models for each region (Table 1). All models were checked for errors and showed low autocorrelation values (range: -0.312 – 0.283) and partial autocorrelation values (range: -0.318 – 0.261) without patterning, thus fulfilling the requirementsfor modeling.[11]

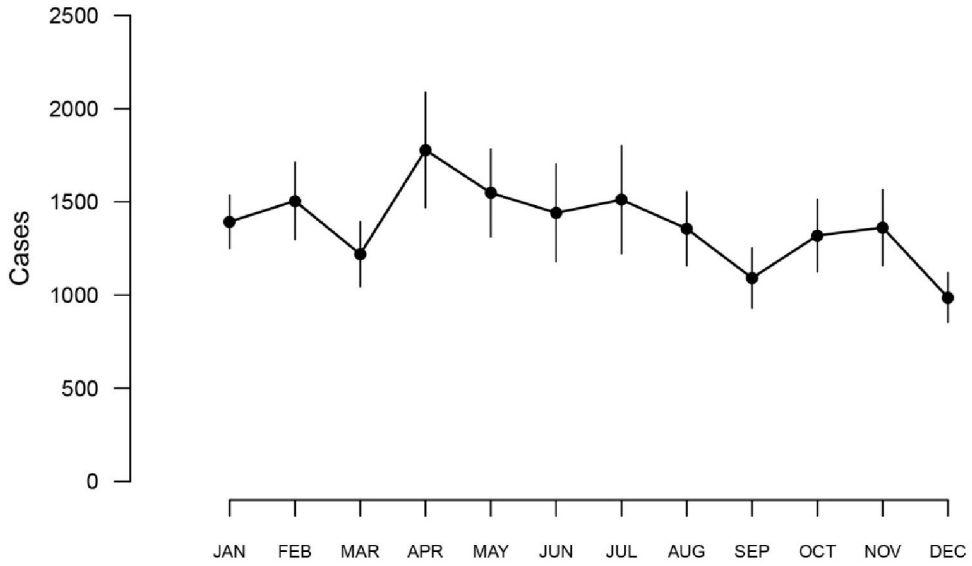

Figure 2 - Monthly mean numberof new TB registered cases in Kazakhstan in 2007-2012
Notes. Error bars represent one standard deviation above and below the mean. (provided in the separate file)

Table 1 - Optimal ARIMA models

Region	**Optimal ARIMA model**	**ACF***	**PACF***
Akmola	SARIMA(2,1,0)(2,0,0)$_{12}$	0.283	-0.283
Aktobe	SARIMA(2,1,1)(1,0,1)$_{12}$	-0.156	-0.160
Almaty	SARIMA(2,0,2)(1,0,1)$_{12}$	0.184	0.178
Atyrau	SARIMA(1,1,1)(1,0,0)$_{12}$	-0.261	-0.210
West-Kazakhstan	SARIMA(2,1,0)(0,0,1)$_{12}$	-0.160	-0.174
Zhambyl	SARIMA(1,1,1)(2,0,0)$_{12}$	0.266	0.251
Karaganda	SARIMA(0,0,0)(0,1,1)$_{12}$	0.238	0.261
Kostanay	SARIMA(2,1,0)(1,0,0)$_{12}$	-0.235	-0.261
Kyzylorda	ARIMA(4,1,2)	-0.205	-0.186
Mangystau	SARIMA(1,1,1)(1,0,0)$_{12}$	-0.181	-0.187
South-Kazakhstan	SARIMA(2,1,1)(2,0,0)$_{12}$	-0.199	-0.223
Pavlodar	SARIMA(1,1,1)(1,0,0)$_{12}$	-0.271	-0.318
North-Kazakhstan	SARIMA(0,1,1)(1,0,1)$_{12}$	-0.280	-0.262
East-Kazakhstan	SARIMA(0,1,1)(2,0,0)$_{12}$	-0.312	-0.313
City of Astana	SARIMA(0,1,1)(1,1,1)$_{12}$	-0.266	-0.273
City of Almaty	SARIMA(1,0,0)(2,0,1)$_{12}$	-0.183	-0.197
Republic of Kazakhstan	SARIMA(0,1,1)(1,0,1)$_{12}$	-0.197	-0.214

Notes. ARIMA = autoregressive integrated moving average; SARIMA = seasonal autoregressive integrated moving average; ACF = autocorrelation function PACF = partial autocorrelation function

*Residual ACF and PACF valueshavingmaximum absolute values.

Performance of forecasting

The predicted and actual values of new registered cases are presented in Table 2. Based on a comparison of percentage errors, Model 1 has the best prediction in June (5.0% error) and the worst in November (22.6% error). Model 2 has the best prediction in September (3.7 error)and the worst in November (17.3% error). Model 3 has the best prediction in July (-1.3% error) and the worst in November (13.2% error). Models consistently overestimated the cases (as compared to actual 2013 case registry) except for July in Model 3. Table 3 shows the model performance assessment estimates. Model 3, the forecasting model built as the sum of regional models, showed the best performance (MAD = 70.3, RMSE = 80, MAPE = 7.3) in the comparison of forecasted and actual values (Table 3).

Table 2 - Predicted and actual number of new TB cases notified by month in 2013

	Model 1		Model 2		Model 3		New TB cases registered
	Predicted cases	PE*	Predicted cases	PE*	Predicted cases	PE*	
January	1166	10.5	1183	12.1	1132	7.3	1055
February	1230	7.2	1239	8.0	1168	1.8	1147
March	1012	10.7	998	9.2	957	4.7	914
April	1480	13.8	1453	11.8	1317	1.3	1300
May	1247	10.3	1270	12.3	1189	5.1	1131
June	1086	5.0	1149	11.1	1139	10.2	1034
July	1262	8.8	1221	5.3	1145	-1.3	1160
August	1163	19.8	1123	15.7	1090	12.3	971
September	944	7.8	908	3.7	946	8.0	876
October	1128	17.5	1111	15.7	1068	11.3	960
November	1182	22.6**	1131	17.3**	1091	13.2**	964
December	863	12.2	820	6.6	852	10.8	769
Year total	13,763	11.2	13,606	10.1	13,094	6.5	12,281
Model to observed difference	1,482		1,325		813		

Notes.ARIMA = autoregressive integrated moving average; SARIMA = seasonal autoregressive integrated moving average
Model 1 = Previous year data SARIMA $(0,0,0)(0,1,0)_{12}$.
Model 2 = National SARIMA model $(0,1,1)(1,0,1)_{12}$.
Model 3 = Sum of regional ARIMA models predictions.
*PE = percentage error
** largest PE

Table 3 - Forecasting measureerror on number of new TB registered cases in Kazakhstan (2013, per month)

	Model 1	Model 2	Model 3
MAD	123.5	110.4	70.3
RMSE	133.5	118.7	80
MAPE	12.2	10.7	7.3

Notes. MAD = mean absolute deviation; RMSE = root-mean-square error; MAPE = mean absolute percentage error.
Model 1 = Previous year data SARIMA $(0,0,0)(0,1,0)_{12}$.
Model 2 = National SARIMA model $(0,1,1)(1,0,1)_{12}$.
Model 3 = Sum of regional ARIMA models predictions.

Discussion

The national ARIMA model and the regional models correctly reflect Kazakhstan's downward trend in TB incidence and showed a better fit compared to the prediction based on historical data. The sum of the regional models is more accurate than the national model in this high MDR-TB burden country. This is expected because more data points were used for the regional models than were used for the national model only.

The national (model 2) and sum of regional models (model 3) predicted greater incidence in 2013 than was recorded in 2013 (2013recorded - 12,281 cases, model 2predicted 13,606 cases, model 3 predicted 13,094cases). This divergence in the two models from the actual observed data may be due to new intensive efforts to combat tuberculosis in the country. One of the six target indicators of the State Program of Health Care Development "Salamatty Kazakhstan" for 2011-2015 was decreasing incidence of tuberculosis.[12]

A limitation of this study was that we used the TB case notification rate and not the actual incidence. However, we believe that there have been no significant change in the identification of cases, the quality of diagnosis, or the reporting of case, and trends in notification is largely reflects trends in incidence. In 2012 the case notification rate was 111 per 100,000 population while the WHO reported the incident rate as 137 per 100,000 population.[1]

The approach of summing the regional models is valid only if incidence rates are sufficiently high, i.e. the disease is not rare. In such cases we may achieve the underlying distributional assumptions of residuals. If the disease is rare, our prediction can include negative values for cases, which are meaningless, and ARIMA models may be inappropriate. In this situation other approaches should be considered.[13]

Despite the fact that univariate time series analysis doesn't account covariates including demographical, economical, behavioral, epidemiological, health system variables and isn't considered as a commonly used approach to justify TB control measures,[14] it can be used as a practical tool in public health practice. First, it's very important to plan drug procurement properly. Issues with drug stock-outs have occurred in many countries[15,16]with implications such as delaying treatment or choosing alternative treatment regime.[17,18] Treatment interruptions, including those due to drug stock-outs, are associated with poor outcomes.[19,20] Second, mismatch between the predicted and actual numbers can be applied as a flag to revise, reinforce or keep up current TB control activities.

The forecasting approach to predict number of new cases using ARIMA models is being implemented in everyday practice of TB system in Kazakhstan for both drug-susceptible and MDR-TB new cases and should be considered for adoption in other high burden areas.

Acknowledgements

The authors thank the National Center for Tuberculosis in Kazakhstan for kindly providing access to the patient population for this study. The authors also thank the Center for Life Sciences of Nazarbayev University for assisting us in securing study funding. The Ministry of Education and Science of the Republic of Kazakhstan financially supported this study.

Ethical Approval

This protocol was reviewed and approved by the Columbia UniversityInstitutional Review Board.

Funding Source: Ministry of Education and Science of the Republic of Kazakhstan

Conflict of Interest: None to declare.

REFERENCES

[1] World Health Organization. Global tuberculosis report 2013.

[2] Schluger NW, El-Bassel N, Hermosilla S, Terlikbayeva A, Darisheva M, Aifah A, et al. Tuberculosis, drug use and HIV infection in Central Asia: an urgent need for attention. Drug Alcohol Depend. 2013 Nov;132Suppl 1:S32-6.

[3] World Health Organization: TB impact measurement. Policy and recommendations for how to assess the epidemiological burden of TB and the impact of TB control. Stop TB policy paper; no. 2. World Health Organization Document 2009, 1-58, WHO/HTM/TB/2009.416.

[4] Mabaera B, Naranbat N, Katamba A, Laticevschi D, Lauritsen JM, Rieder HL. Seasonal variation among tuberculosis suspects in four countries. Int Health. 2009 Sep;1(1):53-60.

[5] Nagayama N, Ohmori M. Seasonality in various forms of tuberculosis. Int J Tuberc Lung Dis. 2006 Oct;10(10):1117-22. PubMed PMID: 17044204.

[6] Kumar V, Singh A, Adhikary M, Daral S, Khokhar A, Singh S. Seasonality of tuberculosis in Delhi, India: a time series analysis. Tuberc Res Treat. 2014;2014:514093.

[7] Willis MD, Winston CA, Heilig CM, Cain KP, Walter ND, Mac Kenzie WR. Seasonality of tuberculosis in the United States, 1993-2008. Clin Infect Dis. 2012 Jun;54(11):1553-60.

[8] Permanasari AE, Rambli DR, Dominic PD. Performance of univariate forecasting on seasonal diseases: the case of tuberculosis. AdvExp Med Biol. 2011;696:171-9.

[9] Moosazadeh M, Nasehi M, Bahrampour A, Khanjani N, Sharafi S, Ahmadi S. Forecasting tuberculosis incidence in iran using box-jenkins models. Iran Red Crescent Med J. 2014 May;16(5):e11779.

[10] Hyndman R.J. and Khandakar Y. (2008) Automatic time series forecasting: The forecast package for R, Journal of Statistical Software, 26(3).

[11] Box GEP and Pierce DA. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models Journal of the American Statistical Association Vol. 65, No. 332 (Dec., 1970), pp. 1509-1526

[12] «Salamatty Kazakhstan» National Healthcare Development Program. http://primeminister.kz/program/about/index/21 (accessed September 12, 2014).

[13] Quddus MA. Time series count data models: an empirical application to traffic accidents. Accid Anal Prev. 2008 Sep;40(5):1732-41.

[14] Dowdy DW, Houben R, Cohen T, Pai M, Cobelens F, Vassall A, et al. Impact and cost-effectiveness of current and future tuberculosis diagnostics: the contribution of modelling. Int J Tuberc Lung Dis. 2014 Sep;18(9):1012–1018

[15] D'Ambrosio L, Dara M, Tadolini M, Centis R, Sotgiu G, van der Werf MJ, et al; European national programme representatives. Tuberculosis elimination: theory and practice in Europe. EurRespir J. 2014 May;43(5):1410-20.

[16] Centers for Disease Control and Prevention (CDC). Interruptions in supplies of second-line antituberculosis drugs--United States, 2005-2012. MMWR Morb Mortal Wkly Rep. 2013 Jan 18;62(2):23-6.

[17] Centers for Disease Control and Prevention (CDC). Impact of a shortage of first-line antituberculosis medication on tuberculosis control - United States, 2012-2013. MMWR Morb Mortal Wkly Rep. 2013 May 24;62(20):398-400.

[18] Pettit AC, Cummins J, Kaltenbach LA, Sterling TR, Warkentin JV. Non-adherence and drug-related interruptions are risk factors for delays in completion of treatment for tuberculosis. Int J Tuberc Lung Dis. 2013 Apr;17(4):486-92.

[19] Podewils LJ, Gler MT, Quelapio MI, Chen MP. Patterns of treatment interruption among patients with multidrug-resistant TB (MDR TB) and association with interim and final treatment outcomes. PLoS One. 2013 Jul 29;8(7):e70064.

[20] Jakubowiak W, Bogorodskaya E, Borisov S, Danilova I, Kourbatova E. Treatment interruptions and duration associated with default among new patients with tuberculosis in six regions of Russia. Int J Infect Dis. 2009 May;13(3):362-8.

Б. Жусупов, S. Hermosilla[b], А. Терликбаева, А. Aifah[b], З. Жумадилов, Т. Абильдаев, Т. Муминов, Р. Исаева

Колумбийский университет, Центр Изучения Глобального Здоровья в Центральной Азии; ул. Луганского, 102, Алматы, 050051, Казахстан;
Колумбийский университет в Нью-Йорке; 722 Запад, 168--я улица № 507, бокс 13, Нью-Йорк 10032, Соединенные Штаты Америки;
Центр наук о жизни «Университета Назарбаева»; ул. Кабанбай батыра, 5, Астана, 010000, Казахстан;
Национальный центр проблем туберкулеза в Казахстане; ул. Бекхожина, 5, Алматы 050059, Казахстан;
Казахстанская ассоциация специалистов по туберкулезу;

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ ПО НОВЫМ СЛУЧАЯМ ТУБЕРКУЛЕЗА В КАЗАХСТАНЕ

Аннотация. Оценить прогнозируемую национальную модель временных рядов по заболеваемости туберкулезом (ТБ), построенную как сумму региональных моделей по сравнению с моделью, основанной только на национальных данных.

Ключевые слова: туберкулез, АРИМА, прогнозирование заболеваемости.

Б. Жусупов, S. Hermosilla[b], А. Терликбаева, А. Aifah[b], З. Жумадилов, Т. Абильдаев,Т. Муминов, Р. Исаева

Колумбия университеті, Орталық Азиядағы ғаламдық денсаулық сақтау оқыту орталығы,Луганск к-сі,102. Алматы, 050051, Қазақстан;
Нью-Йорктегі Колумбия университеті, Нью-Йорк, 10032. Америка Құрама Штаты;
Назарбаев Университетінің «Өмір туралы ғылымдар орталығы», Қабанбай батыр, 5. Астана қ., 010000,Қазақстан;
Қазақстан туберкулез проблемаларының ұлттық орталығы, Бекхожин, 5. Алматы қ., 050059, Қазақстан;
Туберкулез бойынша Қазақстан мамандар қауымдастығы

ҚАЗАҚСТАНДА ТУБЕРКУЛЕЗДІҢ ЖАҢА ЖАҒДАЙЛАРЫ БОЙЫНША УАҚЫТТЫҚ ТІЗБЕКТІ ТАЛДАУ

Аннотация. Ұлттық мәліметтерге сүйенген үлгімен салыстыра отырып, өңірлік үлгі жиынтығы ретінде түзілген туберкулезбен науқастанудың уақыттық тізбегін болжамды ұлттық моделін бағалау.

Тірек сөздер: туберкулез, ауыруды болжау.