การพัฒนาโมเดลพยากรณ์ทางคอมพิวเตอร์เพื่อประเมินสุขภาพชุมชนจากผลตรวจสุขภาพประจำปี

ลิขิต ปรียานนท์[1*] ณภัทร สองทวี[2] May Soe Thu[1] พรรณี พิเดช[3] ประสงค์ แคน้ำ[4] และ
อภิลักษณ์ วรชาติชีวัน[1]

[1]*ภาควิชาเทคนิคการแพทย์ชุมชน คณะเทคนิคการแพทย์ มหาวิทยาลัยมหิดล จังหวัดนครปฐม*
[2]*ภาควิชาเคมีคลินิก คณะเทคนิคการแพทย์ มหาวิทยาลัยมหิดล จังหวัดนครปฐม*
[3]*งานบริการสุขภาพชุมชน คณะเทคนิคการแพทย์ มหาวิทยาลัยมหิดล จังหวัดนครปฐม*
[4]*ศูนย์พัฒนามาตรฐานและการประเมินผลิตภัณฑ์ คณะเทคนิคการแพทย์ มหาวิทยาลัยมหิดล จังหวัดนครปฐม*

บทคัดย่อ

ผลตรวจสุขภาพประจำปีมีการนำมาใช้อย่างกว้างขวางในการคัดกรองและประเมินสุขภาพของชุมชนเพื่อจำแนกผู้ที่มีความเสี่ยงโดยอาศัยค่าอ้างอิงของแต่ละการทดสอบทางห้องปฏิบัติการ ทว่าผลประเมินสุขภาพประจำปีเป็นการตรวจเพียงครั้งเดียวจึงสะท้อนสภาวะทางสุขภาพของบุคคลในชั่วขณะที่รับการตรวจ และเนื่องจากผลตรวจทางสุขภาพสามารถปรับเปลี่ยนได้ตลอดเวลาจากปัจจัยความหลากหลายทางชีวภาพ ทำให้การนำข้อมูลมาใช้เพื่อประเมินสุขภาพชุมชนอาจมีความผิดพลาดหรือคลาดเคลื่อน แม้การตรวจประเมินสุขภาพเป็นระยะ จะทำให้สามารถประเมินสุขภาพชุมชนได้แม่นยำมากขึ้นแต่มีข้อจำกัดในเรื่องการบริหารจัดการ งบประมาณ เวลา และกำลังคน ทำให้เกิดข้อจำกัดในทางปฏิบัติ การศึกษานี้ได้พัฒนาโมเดลทางคอมพิวเตอร์โดยเทคนิค Markov Chain เพื่อช่วยพยากรณ์รูปแบบการกระจายตัวของผลตรวจสุขภาพเพื่อสามารถนำมาประเมินสภาวะสุขภาพและผลการส่งเสริมสุขภาพได้ถูกต้องมากขึ้น ผลการวิเคราะห์พบว่าโมเดลสามารถพยากรณ์ค่าเฉลี่ยของการกระจายตัวในอนาคตได้แม่นยำกว่าเมื่อเทียบกับผลตรวจสุขภาพเพียงครั้งเดียว ช่วยให้นักเทคนิคการแพทย์หรือผู้เกี่ยวข้องในการส่งเสริมสุขภาพสามารถใช้ข้อมูลในการประเมินสุขภาพได้อย่างเหมาะสม แม้จะมีข้อมูลจำกัด การศึกษานี้เป็นแนวทางใหม่ในการนำข้อมูลสุขภาพที่มีอยู่ในห้องปฏิบัติการเทคนิคการแพทย์มาประยุกต์ใช้เพื่อประโยชน์ในการดูแลสุขภาพชุมชน

คำสำคัญ: การส่งเสริมสุขภาพ สุขภาพชุมชน การดูแลสุขภาพระดับปฐมภูมิ ห่วงโซ่มาร์คอฟ

***ผู้รับผิดชอบบทความ E-mail address:** likit.pre@mahidol.edu

รับบทความ: 15 ตุลาคม 2562 แก้ไขบทความ: 8 มกราคม 2563 รับตีพิมพ์บทความ: 15 มกราคม 2563

J Med Tech Assoc Thailand, Vol. 47 No. 3, December 2019 *Original Article*

Development of Predictive Computer Model for Assessment of Community Health from Annual Health Check-up

Likit Preeyanon[1*], Napat Songtawee[2], May Soe Thu[1], Phanee Pidetcha[3],
Prasong Khaenam[4] and Apilak Worachartcheewan[1]

[1]Department of Community Medical Technology, Faculty of Medical Technology, Mahidol University,
Nakhon Pathom Province, Thailand
[2]Department of Clinical Chemistry, Faculty of Medical Technology, Mahidol University,
Nakhon Pathom Province, Thailand
[3]Community Health Care Service Unit, Faculty of Medical Technology, Mahidol University,
Nakhon Pathom Province, Thailand
[4]Center for Standardization and Product Validation, Faculty of Medical Technology, Mahidol University
Nakhon Pathom Province, Thailand

Abstract

Health check-up program has been widely used to assess community health status and to identify people who are at risk according to a significant range of laboratory parameters. The laboratory results from a single health check-up is, however, a snapshot of individual health at a certain time point in which fluctuation due to biological variation resulting in the unreliable measurement of the community health. Laboratory result data collected from multiple health check-ups could be used to improve the assessment of community health. However, it is not practical to perform frequent health check-ups due to limited resources such as workforce, time and budget. In this study, the Markov Chain computational model was developed to predict the average distribution of health parameters using data from a single check-up. The results demonstrated that the established model gave a better prediction of the average distribution from three consecutive check-ups than a single check-up. Therefore, this could be used by Medical Technologists and healthcare promoters to predict health status to gain better insight of community health. This study also revealed the innovative use of longitudinal health check-up data in health promotion.

Keywords: Health promotion, Community health, Primary health care, Markov Chain

---

***Corresponding author E-mail address:** likit.pre@mahidol.edu

Received: 15 October 2019 Revised: 8 January 2020 Accepted: 15 January 2020

Introduction

Population or community health can be defined as the distribution of health outcomes of a group of individuals(1). Many Health professionals, researchers, policymakers and volunteers have actively participated in improving community health under the management of governmental and non-governmental organizations, local authorities and private agencies(2). Recently in Thailand, the healthcare services have been appointed to focus on primary health care to promote individual and community health(3). A health check-up and screening program has been used nation-wide to assess health status of the community and to identify healthy people who are at risk of health problems or conditions such as diabetes, cardiovascular disease and metabolic syndrome.(4-8). Although individual and community health are intertwined and context-dependent(9), a number of people with health issue risk are considered the key performance indicators (KPIs) that need to be monitored and minimized by local health promoters(10).

In practice, assessing the community health using data from the health check-up has shown some drawbacks. The health check-up result parameters of individual fluctuate in a period of time due to biological variation, time of assessment and other unknown factors(11-15). These fluctuations may occur between annual health check-ups and affect the distribution of the health parameters in the community. While an average of results gathered from health check-ups could provide a more reliable assessment for individual and community health, it is not always feasible to perform multiple health check-ups to cover all people in the community due to limited time, budget and human resources.

A computational model that could predict average results of the laboratory test using data from single health check-up would be useful for assessment of the community health. However, the average distribution is difficult to predict because it depends on the results from each individual which fluctuates within a certain period of time. In healthcare monitoring, the multistate Markov model has been used successfully to predict health outcome and progression of disease transitioning among defined illnesses such as diabetes, cardiovascular disease, cancer and metabolic syndrome(16-20). For the health check-up, the change of the individual results between health check-ups reflects a variation of biological state that could be modeled by the Markov process. This study demonstrates the use of the Markov model to predict the average distribution of laboratory test results from a single health check-up that could serve as a more reliable measurement of the community health.

Materials and Methods

Data

The training dataset for the construction of transition matrix and test dataset were obtained from the health data warehouse at the Faculty of Medical Technology, Mahidol University, Thailand. The datasets included data from the health check-up customers age ranges between 35-55 years participated in the annual health check-up service from 2011-2015 in Bangkok and its vicinity. The train dataset included the customers who participated in the check-up from 2011-2015 consecutively and the test dataset included a group of customers who participated in the check-up services from 2012-2014 consecutively. The train and test datasets have no common customers. (Table 1)

Table 1 The number of customers in the train and test datasets used to develop and evaluate the model of each laboratory test

Test	Training Dataset	Testing Dataset
Fasting Blood Sugar (FBS)	2738	3795
Triglyceride (TG)	2752	3778
Low-density lipoprotein cholesterol (LDL-C)	2555	3652
High-density lipoprotein cholesterol (male) (HDL-C)	2710	1427
High-density lipoprotein cholesterol (female) (HDL-C)	1607	2355
Total Cholesterol (CHOL)	2783	3831

Mean individual standard deviation

To construct the transition matrix for the Markov model, the results from a test were divided into bins of equal width. The standard variation (SD) of an individual's results from 2011-2015 were calculated to set the width of the bin that represents a biological state. All standard deviations were then averaged to find the mean SD of the group. Then 2.0 x SD or 2SD was used as the width of the bin. For example, the mean SD of the fasting blood sugar test is 6.04 mg/dL; therefore, the width of each bin is equal to 12.0 mg/dL. The 2SD was used based on the assumption that the results of the individual at a certain biological state could fluctuate within this range. Thus, the change of the result greater than 2SD signifies the change of the biological state. The mean SD and the bin width of each test are shown in Table 2.

Table 2 The mean individual standard deviation and the bin width of each test

Test	Mean Individual Standard Deviation (SD)	Bin Width (2SD)
FBS	6.04	12.0
HDL-C (male)	5.52	11.0
HDL-C (female)	5.92	11.6
TG	28.57	57.0
LDL-C	14.63	29.2
CHOL	15.54	31.0

Transition probability matrix construction

The data from all individuals who participated in the health check-up program consecutively from 2011-2015 were used to construct the transition matrix. Using results from the first check-up, the difference of each test result and the cutoff value was first calculated and the difference was then categorized into a bin whose width was obtained from the mean SD as described above. Each bin represents the reversible state, which is a range of the test result of an individual. The annualized probability of the transition was obtained from the frequency of transition of the value to the same bin or a different bin in the following check-up.

Prediction of average results distribution

To predict the average results distribution, the initial frequency matrix was calculated from the frequency in each bin as described in the method section. The initial matrix was then multiplied with the transition probability matrix to calculate the frequency of the results in each bin in the following n year according to the formula:

$$\chi^{(n+1)} = \chi^{n}P$$

where P is the transition probability matrix and χ is the frequency matrix. The frequency in each bin from n years were averaged to derive a predicted frequency for each range of the test results.

Ethical declaration

Ethical approval was delivered by the ethics committee at Mahidol University, Thailand. The protocol number is MU-CIRB 2017/027.2401 and certificate of exemption (COE) number is MU-CIRB 2017/001.2701. MU-CIRB is in full compliance with international guidelines for human research protection such as the declaration of Helsinki, the Balmont report, CIOMS guidelines and the international conference on harmonization in good clinical practice (ICH-GCP).

Results

From the fasting blood sugar (FBS) results shown in Fig. 1, the distribution of the FBS results from three years varied from 66.0 to 113.0 mg/dL. According to these data, the health promoters could compare the data from 2012 and 2013 to evaluate the community health status. These could be interpreted that a number of people with the borderline FBS level (90-101 mg/dL) increased consecutively. However, data collected during 2013 and 2014 showed a decrease in number of people with the borderline FBS level. These results demonstrated that the number of people with a particular FBS level could fluctuate within a three-year period of time leading to different interpretation and prediction. As a result, the health promoters should always be cautious when assessing the health status using data from a single health check-up.

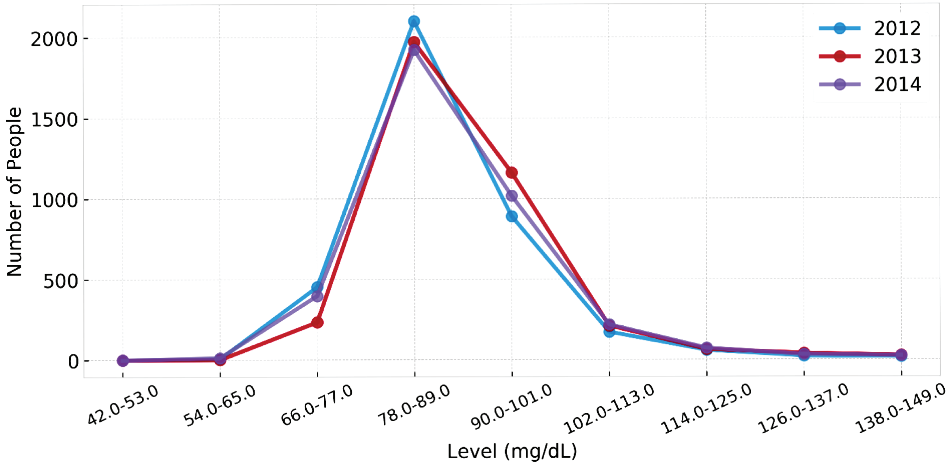

Fig. 1 The distribution of FBS levels between 42.0-149.0 mg/dL from the test dataset with the bin width equal to 12.0 mg/dL

Due to lack of a true distribution of the FBS results for the community, the average distribution is derived by averaging the number of people in each bin from 2012-2014. Fig. 2 showed the average distribution and the distribution of FBS levels from 2012-2014. Compared with the average distribution, the number of people with the borderline FBS level in 2012 appeared to be underestimated whereas the number of people with the borderline FBS level in 2013 appeared to be overestimated. On the other hand, the distribution in 2014 was highly similar to the average. The average distribution was used to represent the distribution of FBS levels for this community and serves as a benchmark for prediction.

In general, a series of health check-ups must be performed to obtain a reliable average

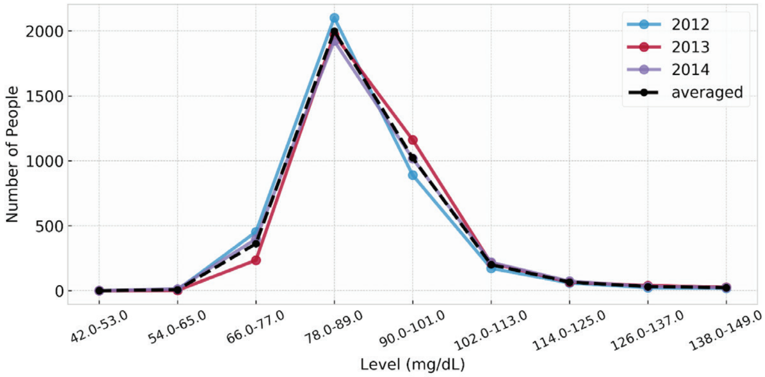

Fig. 2 The average distribution along with the distribution of the FBS levels between 42.0-149.0 mg/dL from the test dataset

distribution. Unfortunately, the community-wide health check-up requires a substantial amount of budget, time, staff and other resources which may not be available on a yearly basis. The data from the latest health check-up was often used to assess the health status of the community which could be misleading as shown before. To address this problem, the Markov Chain (MC) method was used to develop a model that predicts the average distribution of data from one check-up. The transition matrix was constructed from the train dataset as described in the method section. The predicted distribution of the test dataset is shown in Fig. 3. The correlation coefficient (R^2) and root mean squared error

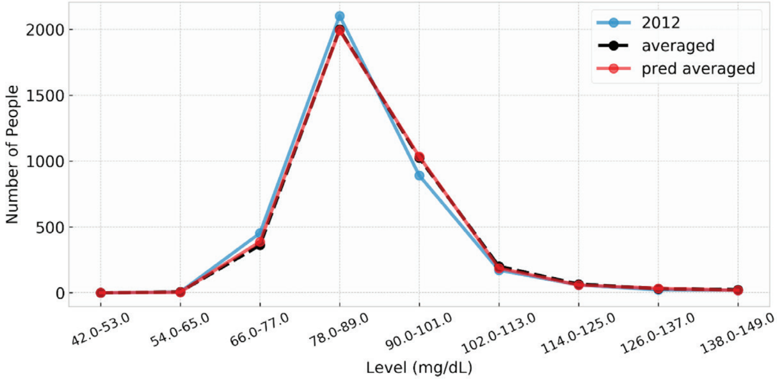

Fig. 3 The predicted distribution, the 3-year average distribution and the distribution of the FBS levels in 2012 from the test dataset. The predicted distribution calculated from the MC model using only the data from 2012 was highly similar to the 3-year average distribution, which could help health promoters evaluate the health status of a community even without subsequent check-ups

(RMSE) between the 3-year average distribution, the predicted distribution, and the distribution of FBS in 2012 are shown in Table 3. The scatter plot of the number of people between the 3-year average distribution, the predicted distribution and the distribution of FBS levels in 2012 are shown in Fig. 4A and 4B. Although the R^2 values from the prediction and the data from 2012 were similar, the RMSE from the prediction was much lower suggesting that the model could estimate the average distribution more accurately than the data from 2012.

Table 3 The correlation coefficient (R^2) and the root mean squared error (RMSE) from the FBS levels between 42.0-149.0 mg/dL.

Method	Correlation Coefficient (R^2)	Root Mean Squared Error (RMSE)
Predicted mean	0.99	11.00
Year 2012	0.98	64.43

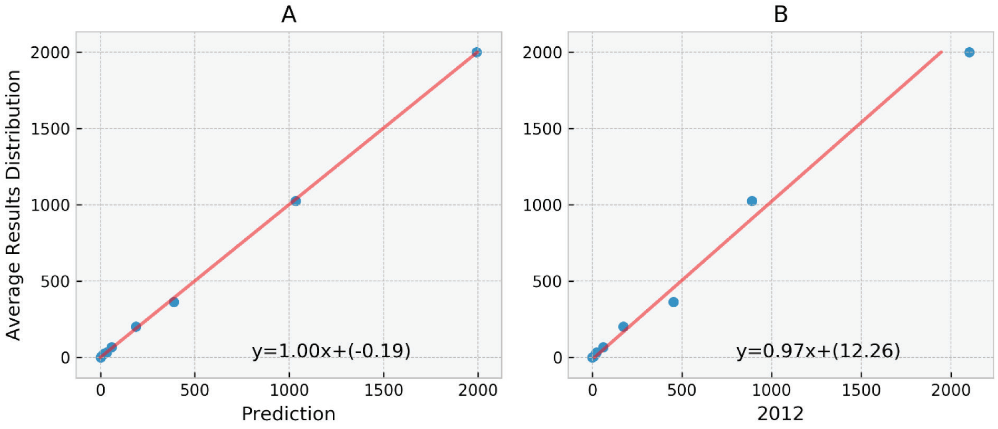

Fig. 4 The scatter plot and the fit curve of the number of people in each bin of FBS levels between 42.0-149.0 mg/dL between the 3-year average distribution and A) the predicted distribution and B) the distribution of FBS levels in 2012

For the diabetes mellitus (DM) screening program, people with the FBS level between 100-125 mg/dL could be classified in a risk group, which may need a proper intervention to prevent the development of diabetes mellitus in the future. The results in Table 4 showed that the number of people in this group increased annually in 2012 to 2014.

The increase rate could reflect the ineffectiveness of the health promotion campaign. However, the prediction suggested that the number of people in this group was underestimated in 2012, which was realized in the following years when the number of people in a risk group went up to 278 and 292 respectively (19.3% and 25.3% increase). Thus, one must be careful when use the data from 2012 as a baseline for community health evaluation before judging the effectiveness of health promotion.

The model was further developed for the triglyceride (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C) and total cholesterol (CHOL) test. For HDL-C, the model was built for male and female because of the different cutoff in male (<40 mg/dL) and female (<50 mg/dL). The results shown in Fig. 5-9 and Table 5 suggested that the model could be used to predict the average results distribution with higher accuracy when compared to the results distribution in 2012.

Table 4 The number of people in the risk group from 2012-2014, an average and the prediction

Results	2012	2013	2014	3-year Average	Prediction
102.0-125.0	233	278	292	269	248

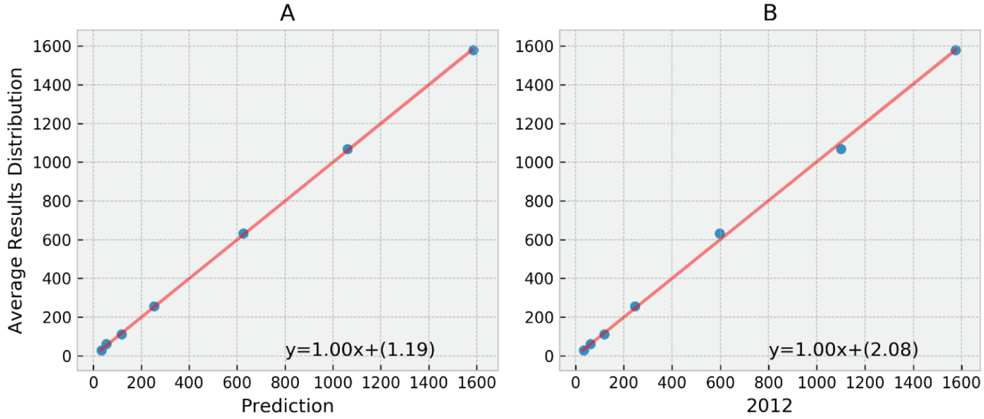

Fig. 5 The scatter plot and the fit curve of the number of people from each bin of TG levels between 14.0-413.0 mg/dL from the 3-year average distribution and A) the predicted distribution and B) the distribution of TG levels in 2012

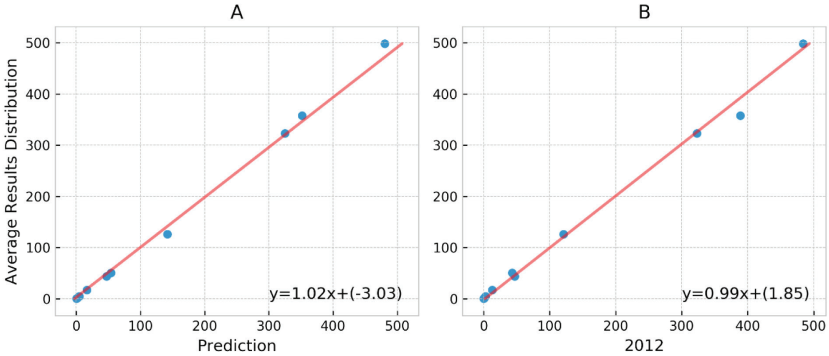

Fig. 6 The scatter plot and the fit curve of the number of people from each bin of male HDL-C levels between 9.0-130.0 mg/dL from the 3-year average distribution and A) the predicted distribution and B) the distribution of male HDL-C levels in 2012

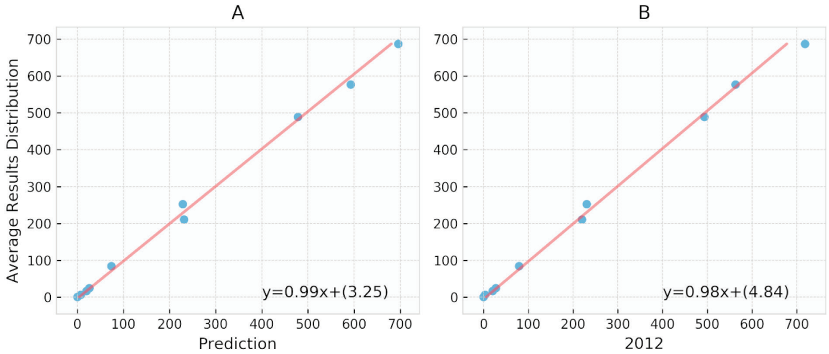

Fig. 7 The scatter plot and the fit curve of the number of people from each bin of female HDL-C levels between 20.4-135.4 mg/dL from the 3-year average distribution and A) the predicted distribution and B) the distribution of female HDL-C levels in 2012

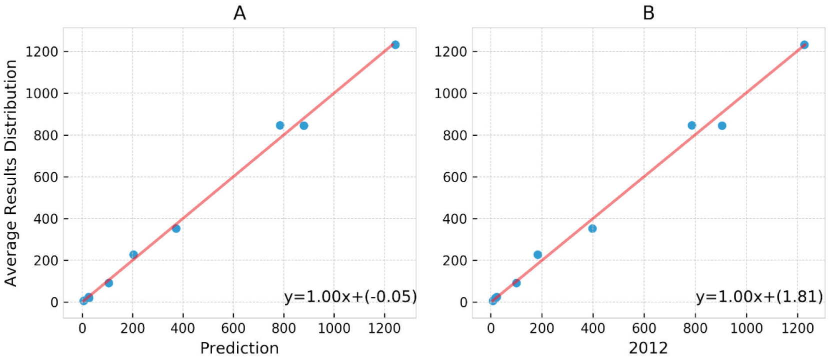

Fig. 8 The scatter plot and the fit curve of the number of people from each bin of LDL-C levels between 26.0-288.8 mg/dL from the 3-year average distribution and A) the predicted distribution and B) the distribution of LDL-C levels in 2012

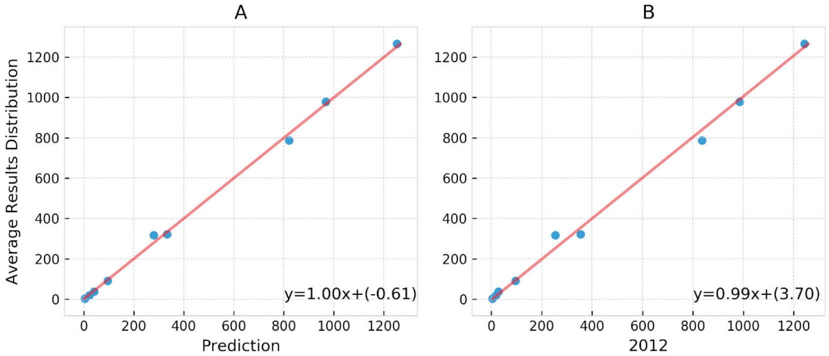

Fig. 9 The scatter plot and the fit curve of the number of people from each bin of CHOL levels between 74.0-353.0 mg/dL from the 3-year average distribution and A) the predicted distribution and B) the distribution CHOL levels in 2012

Table 5 The results from the evaluation of the predicted average results distribution of TG, HDL-C, LDL-C, and CHOL

Test	Cutoff	Result Range	Correlation Coefficient (R^2)		Root Mean Squared Error (RMSE)	
			Prediction	2012	Prediction	2012
TG	150 mg/dL	14.0-413.0 mg/dL	0.99	0.99	6.44	18.78
HDL-C (male)	40 mg/dL	9.0-130.0 mg/dL	0.99	0.98	7.61	10.83
HDL-C (female)	50 mg/dL	20.4-135.4 mg/dL	0.99	0.99	12.61	13.60
LDL-C	130 mg/dL	26.0-288.8 mg/dL	0.99	0.99	26.62	35.07
CHOL	200 mg/dL	74.0-353.0 mg/dL	0.99	0.99	18.44	30.15

Discussion

The model built from the Markov Chain method using the longitudinal health check-up data estimates the average distribution of the laboratory tests with more accuracy than data from a single check-up as suggested by the RMSE. Although the model was built using mainly the data from people who live or work in Bangkok and its vicinity, the model could be used in other parts of the country because a great number of people who live and work in Bangkok are from all parts of the country. In case the data is available, the local health promoters are suggested to build the model using the local health data because the difference among communities may be significant.

Furthermore, the health data used in this study were from people age ranged between 35-55 years old; therefore, the model might not be applicable to other age groups. We used this group mainly because a majority of the customers of the health check-up service was an active worker of an organization. More data have to be collected for people under 35 and over 55 to be able to apply the method to a wider group of people. Note that the biological variation between each age group in each community might be significantly different, so the age group has to be adjusted properly according to the health data from the community to get the most accurate results.

Conclusion

Data from a health check-up are useful for monitoring and evaluating the health status of the community. However, the results from a single health check-up are not reliable due to the fluctuation of individual results. The Markov chain method was then applied in this study to build a model that could predict the average distribution of the health check-up results, which provides more accurate measurement of the community health. This study also highlights the great value and innovative use of the longitudinal health check-up data in health analytics that will play an important role in data-driven health promotion in the future. An organization or the government should therefore encourage and provide support for an individual to participate in a regular health check-up and use of health data in the community.

References

1. Kindig D, Stoddart G. What is population health? Am J Public health 2003; 93: 380-3.
2. Wilson MG, Lavis JN, Guta A. Community-based organizations in the health sector: A scoping review. Health Res Policy Syst 2012; 10: 36.
3. Ministry of Public Health. Twenty-Year National Strategic Plan for Public Health (2017-2036) First Revision 2018. 1 ed. Bangkok: Strategy and Planning Division, Office of the Permanent Secretary, Ministry of Public Health, Thailand; 2018.
4. Teerawattananon Y, Kingkaew P, Koopitakkajorn T, *et al.* Development of a health screening package under the universal health coverage: the role of health technology assessment. Health Econ 2016; 25: 162-78.
5. Charuruks N, Milintagas A, Watanaboonyoungcharoen P, *et al.* Annual laboratory checkup: early signs of health problems in young and middle-age adults. Southeast Asian J Trop Med Public Health 2005; 36: 769.
6. Boonlert W, Kost GJ, Jiraviriyakul A, *et al.* Point-of-care testing on a mobile medical unit in Northern Thailand: screening for hyperglycemia, hyperlipidemia, and Thalassemia trait. Point of Care 2006; 5: 164-7.

7. Uaratanawong S, Suraamornkul S, Angkeaw S, *et al.* Prevalence of hyperuricemia in Bangkok population. Clin Rheumatol 2011; 30: 887-93.
8. Rose G. Sick individuals and sick populations. Int J Epidemiol 2001; 30: 427-32.
9. Arah OA. On the relationship between individual and population health. Med Health Care Philos 2009; 12: 235-44.
10. Ministry of Public Health. The Detail Report of the Key Performance Indicators of the Ministry of Public Health for the Fiscal Year 2018. 1 ed. Bangkok: Strategy and Planning Division, Office of the Permanent Secretary, Ministry of Public Health, Thailand; 2018.
11. Garde AH, Hansen ÅM, Skovgaard LT, *et al.* Seasonal and biological variation of blood concentrations of total cholesterol, dehydroepiandrosterone sulfate, hemoglobin A1c, IgA, prolactin, and free testosterone in healthy women. Clin Chem 2000; 46: 551-9.
12. Widjaja A, Morris RJ, Levy JC, *et al.* Within-and between-subject variation in commonly measured anthropometric and biochemical variables. Clin Chem 1999; 45: 561-6.
13. Rotterdam E, Katan M, Knuiman J. Importance of time interval between repeated measurements of total or high-density lipoprotein cholesterol when estimating an individual's baseline concentrations. Clin Chem 1987; 33: 1913-5.
14. Carlsen S, Petersen PH, Skeie S, *et al.* Within-subject biological variation of glucose and HbA1c in healthy persons and in type 1 diabetes patients. Clin Chem Lab Med 2011; 49: 1501-7.
15. Omar F, van der Watt GF, Pillay TS. Reference change values: how useful are they? J Clin Pathol 2008; 61: 426-7.
16. Jia X, Chen Q, Wu P, *et al.* Dynamic development of metabolic syndrome and its risk prediction in Chinese population: a longitudinal study using Markov model. Diabetol Metab Syndr. 2018; 10: 24.
17. Tang X, Liu Q. Prediction of the development of metabolic syndrome by the Markov model based on a longitudinal study in Dalian City. BMC Public Health 2018; 18: 707.
18. Caro JJ, O'Brien JA, Hollenbeak CS, *et al.* Economic burden and risk of cardiovascular disease and diabetes in patients with different cardiometabolic risk profiles. Value in Health 2007; 10: S12-S20.
19. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate Markov models for disease progression with classification error. J Royal Stat Soc 2003; 52: 193-209.
20. Biritwum R, Odoom S. Application of Markov process modelling to health status switching behaviour of infants. Int J Epidemiol 1995; 24: 177-82.