

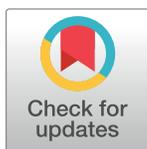
RESEARCH ARTICLE

Prediction of blood test values under different lifestyle scenarios using time-series electronic health record

Takanori Hasegawa^{1*}, Rui Yamaguchi², Masanori Kakuta², Kaori Sawada³, Kenichi Kawatani⁴, Koichi Murashita⁴, Shigeyuki Nakaji³, Seiya Imoto¹

1 Health Intelligence Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan, **2** Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan, **3** Department of Social Medicine, Graduate School of Medicine, Hirosaki University, Hirosaki, Aomori, Japan, **4** COI Research Initiatives Organization, Hirosaki University, Hirosaki, Aomori, Japan

* t-hasegw@ims.u-tokyo.ac.jp



Abstract

Owing to increasing medical expenses, researchers have attempted to detect clinical signs and preventive measures of diseases using electronic health record (EHR). In particular, time-series EHRs collected by periodic medical check-up enable us to clarify the relevance among check-up results and individual environmental factors such as lifestyle. However, usually such time-series data have many missing observations and some results are strongly correlated to each other. These problems make the analysis difficult and there exists strong demand to detect clinical findings beyond them. We focus on blood test values in medical check-up results and apply a time-series analysis methodology using a state space model. It can infer the internal medical states emerged in blood test values and handle missing observations. The estimated models enable us to predict one's blood test values under specified condition and predict the effect of intervention, such as changes of body composition and lifestyle. We use time-series data of EHRs periodically collected in the Hirosaki cohort study in Japan and elucidate the effect of 17 environmental factors to 38 blood test values in elderly people. Using the estimated model, we then simulate and compare time-transitions of participant's blood test values under several lifestyle scenarios. It visualizes the impact of lifestyle changes for the prevention of diseases. Finally, we exemplify that prediction errors under participant's actual lifestyle can be partially explained by genetic variations, and some of their effects have not been investigated by traditional association studies.

OPEN ACCESS

Citation: Hasegawa T, Yamaguchi R, Kakuta M, Sawada K, Kawatani K, Murashita K, et al. (2020) Prediction of blood test values under different lifestyle scenarios using time-series electronic health record. PLoS ONE 15(3): e0230172. <https://doi.org/10.1371/journal.pone.0230172>

Editor: Peyman Rezaei-Hachesu, Tabriz University of Medical Sciences, ISLAMIC REPUBLIC OF IRAN

Received: March 18, 2019

Accepted: February 24, 2020

Published: March 20, 2020

Copyright: © 2020 Hasegawa 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 Statement: Data cannot be shared publicly because of the ethical concerns. Data are available from the Hirosaki University COI Program Institutional Data Access / Ethics Committee (contact via e-mail: coi@hirosaki-u.ac.jp) for researchers who meet the criteria for access to confidential data. Researchers need to be approved by conducting a research ethics review at the organization to which the researchers using the data.

Introduction

Recently, the continuously increasing cost of medical care has received significant attention. The center of the idea aimed at curbing this trend is using electronic health records (EHRs) to detect signs and preventive measures of diseases such as diabetes. EHRs consist of patient's medical information, e.g., demographics, symptoms, and blood test values. Thus, while the

Funding: SI and SN received the Center of Innovation Program from Japan Science and Technology Agency (<https://www.jst.go.jp/EN/>) TH received Grant-in-Aid for Young Scientists (B) Grant Number 17K12647 from Japan Society for the Promotion of Science (<https://www.jsps.go.jp/english/>).

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

essential purpose of EHR is to document the history of patient care for reimbursement, the accumulated EHRs have been gathering a great deal of attention for advanced purposes. For example, the analysis of EHRs enables us to facilitate clinical decision support, to predict condition-specific clinical processes, and to improve clinical treatment plans beyond traditional clinical encounters. In particular, periodic time-series EHRs collected in, *e.g.*, annual medical check-ups, can clarify trends of changes in check-up results depending on lifestyle and social status. However, such time-series data have many missing observations and some results are strongly correlated to each other. These problems make the analysis difficult.

Traditionally, time-aware models, *e.g.*, survival modeling, are utilized when we focus on the time-to-event to evaluate the effect of treatments or occurrence of events for some endpoints, *e.g.*, the mortality of cardiovascular diseases. On the other hand, there exist statistical techniques, which utilize statistical models to mathematically formalize the generative process of the data under assumptions. They can elucidate hidden mechanism of the target system and also incorporate the effect of time such as dynamic Bayesian networks (DBNs) [1, 2] and the state space models (SSM) [3–5]. For example, Nachimuthu *et al.* [6] used DBNs to model temporal relationships between insulin and glucose homeostasis and predicted the future glucose levels of a patient admitted in an ICU. Sandri *et al.* [7] and Gatti *et al.* [8] also used DBNs with the restrictions on the causal structure to model organ failure. Similarly, Peelen *et al.* [9] used a hierarchical Bayesian strategy for modeling organ failure. These researches constructed statistical models for the analysis of target medical incident and successfully obtained clinical findings. Thus, we design statistical models that can represent relevance among check-up results, lifestyle changes, and social status depending on time.

In this paper, we focus on blood test values in EHRs and inference of the effect to blood test values by the changes of body composition values such as Body Mass Index (BMI), lifestyle, and social status. For this purpose, we apply a statistical approach using the SSM, which has successful applications in a wide range of fields [10–14]. Our proposed model is designed to infer hidden state variables, which summarize the internal (unobserved) medical states emerged in blood test values, at each time-point for each person. The model realizes the estimation of the effect of body composition values, lifestyles, and social status to blood test values. For the extraction of meaningful relationships, we suppress the sparse constraint to regulatory effect matrices and infer their parameter values by maximization of the L1 regularized likelihood. To this end, we developed a new algorithm to obtain active sets of parameters and estimate a maximizer of the L1 regularized likelihood using the EM algorithm. The proposed approach is applied to EHR obtained in the Hirosaki cohort study in Japan, which is a free participation medical check-up cohort including men and women, aged from 12 to 91 obtained from 2005 to 2015.

Materials and methods

Periodically collected time-series EHR

Here, we consider that EHR is a set of clinical records with date, and it includes each diagnosis or check-up result. Thus, clinical records obtained at different dates, of which individuals are the same, can be integrated to refer to participant information. Especially, in this study, we consider that EHRs consist of collected certain items at regular interval through cohort studies and regular medical check-ups. Although a predetermined interval of visits may exist for each person's condition, *e.g.*, annual medical check-ups are held at one-year intervals and pre-diabetes patients are advised to check their blood test values every few months, many participants do not actually follow these recommendations. Thus, the most striking characteristic of EHR is the irregularity of participants visits.

Let $y_{k,t}$ be q -dimensional observed blood test values of the k th ($k = 1, \dots, K$) participant at the t th ($t = 1, \dots, T_k$) time point (we here consider the data of the first time of visit as $y_{k,1}$), K is the number of participants, and T_k is the final time of visit of the k th participant. We have a set of observation data $Y = \{Y_1, \dots, Y_K\}$, where $Y_k = \{y_{k,1}, \dots, y_{k,T_k}\}$ and the number of observations of the k th participant is not necessarily T_k because the k th participant might not visit at the t th ($1 < t < T_k$) time-point.

Linear state space model for EHR

The state space model (SSM) has been widely used in time-series analysis and it consists of a state transition equation and an observation equation. The state transition equation models the process of the hidden state variables and the observation equation links the observations to these underlying states. The advantage of SSM is the scalability of the design of the model. In this research, we have three motivations; (a) inference of the effect of environmental factors such as lifestyles to blood test values, (b) prediction of one’s blood test values under certain environmental factors, and (c) estimation of genetic effects as the explanatory factors for prediction errors of blood test values. Moreover, there are two major problems; (i) there are strongly correlated blood test values and (ii) EHR has several missing observations. To overcome these problems, we apply the state space model.

Let $x_{k,t}$ and $z_{k,t}$ be a series of p -dimensional vectors containing hidden variables representing the internal medical states emerged in blood test values and m -dimensional vector containing the values representing environmental factors such as body composition values of the k th ($k = 1, \dots, K$) participant at the t th ($t = 1, \dots, T_k$) time point, respectively. Then, we consider a state space model for EHR represented by

$$x_{k,t} = Ax_{k,t-1} + Gz_{k,t-1} + v_{k,t}, \tag{1}$$

$$y_{k,t} = Hx_{k,t} + w_{k,t}, \tag{2}$$

where $A = (a_1, \dots, a_p)'$ is a $p \times p$ regulatory matrix, $a_i = (a_{i,1}, \dots, a_{i,p})'$ ($i = 1, \dots, p$) is a p -dimensional vector including regulatory effects on the i th hidden variable by other ones, $G = (g_1, \dots, g_p)'$ is a $p \times m$ regulatory matrix, $g_i = (g_{i,1}, \dots, g_{i,m})'$ is an m -dimensional vector representing their regulatory effects on the i th hidden variable by $z_{k,t-1}$, $H = (h_1, \dots, h_q)'$ is a $q \times p$ map matrix, $h_i = (h_{i,1}, \dots, h_{i,p})'$ ($i = 1, \dots, q$) is a p -dimensional vector that maps hidden variables on the i th element of $y_{k,t}$, $v_{k,t} \in R^p$ is a vector of system noise, and $w_{k,t} \in R^q$ is a vector of observational noise. Here, diagonal elements of A are restricted to less than 0.8 and we set system noise as $v_{k,t} \sim N_p(\mathbf{0}_p, Q)$ and observation noise as $w_{k,t} \sim N_q(\mathbf{0}_q, R)$, where Q and R are $p \times p$ and $q \times q$ diagonal matrices. The initial state vector $x_{k,0}$ is assumed to be a Gaussian random vector with mean vector $\mu_{k,0}$ and covariance matrix Σ_0 , i.e., $x_{k,0} \sim N_p(\mu_{k,0}, \Sigma_0)$. Because the relationship between observation variables $y_{k,t}$ can be represented by A and H [15], we can estimate the effect from the lifestyle $z_{k,t}$ to the observation variables $y_{k,t}$ with considering the effects between observation variables $y_{k,t}$ using the proposed state space model. Note that we consider that A , G and H should be sparse matrices.

The SSM consisting of Eqs (1) and (2) represents a process how observation values $y_{k,t}$ are generated from an unobserved dynamic system through a latent state vector $x_{k,t}$. Our aim is to estimate posterior distributions of state vectors $x_{k,t}$ ($t = 1, \dots, T_k$) given a set of observation data $Y_k = \{y_{k,1}, \dots, y_{k,T_k}\}$ in Bayesian-statistics manner. The estimation is carried out sequentially from $t = 1$ to $t = T_k$ by updating the posterior distribution of $x_{k,t}$ according to the dynamics of Eq (1) followed by addition of information of $y_{k,t}$ to the set of observation data according to Eq (2) one after another. The modeling and estimation approach can

naturally handle missing observations in time-series data without external imputations by the following mechanism; if the observation at time t' , $\mathbf{y}_{k,t'}$ ($1 < t' < T$), is a missing case, the update of posterior distribution of $\mathbf{x}_{k,t'}$ is done according to Eq (1) only and the step to add information of $\mathbf{y}_{k,t'}$ is just skipped; then the estimation process from $t = t' + 1$ is continued without inconsistency.

Maximum likelihood estimation using the EM algorithm with L1 regularization

We applied L1 regularization to select effective sets of elements for A , G , and H . Let $\{Y, X\} = \{Y_1, \dots, Y_K, X_1, \dots, X_K\}$ be the complete data set, where $X_k = \{\mathbf{x}_{k,0}, \dots, \mathbf{x}_{k,T_k}\}$ is the set of state variables. Here, $\mathbf{y}_{k,t}$ can be unobserved when the k th individual did not participate the check-up at time t , but all $\mathbf{x}_{k,t}$ are estimated. Furthermore, let the probability densities $P(\mathbf{x}_{k,0})$ and $P(\mathbf{x}_{k,t}|\mathbf{x}_{k,t-1})$ be the p -dimensional Gaussian distributions $N_p(\boldsymbol{\mu}_{k,0}, \Sigma_0)$ and $N_p(F\mathbf{x}_{k,t-1} + G\mathbf{z}_{k,t-1}, Q)$, respectively, and $P(\mathbf{y}_{k,t}|\mathbf{x}_{k,t})$ be the q -dimensional Gaussian distribution $N_q(\mathbf{x}_{k,t}, R)$. Then joint likelihood for the complete data set is given by

$$P(Y, X; \boldsymbol{\theta}) = \prod_{k=1}^K P(\mathbf{x}_{k,0}) \prod_{t \in \mathcal{T}_k} P(\mathbf{x}_{k,t}|\mathbf{x}_{k,t-1}) \prod_{t \in \mathcal{T}_{k,obs}} P(\mathbf{y}_{k,t}|\mathbf{x}_{k,t}), \tag{3}$$

where $\boldsymbol{\theta} = \{A, G, Q, H, R, \boldsymbol{\mu}_{1,0}, \dots, \boldsymbol{\mu}_{K,0}\}$, and \mathcal{T}_k and $\mathcal{T}_{k,obs}$ ($\mathcal{T}_{k,obs} \in \mathcal{T}_k$) are sets of all time points and the observed time points of the k th participant, respectively. In this study, we used the Expectation-Maximization (EM) algorithm [16] to search for the parameter vector $\boldsymbol{\theta}$ that maximizes $P(Y; \boldsymbol{\theta})$ under L1 regularization. The L1 regularized log-likelihood is given by

$$\begin{aligned} \log \int \prod_{k=1}^K P(\mathbf{x}_{k,0}) \prod_{t \in \mathcal{T}_k} P(\mathbf{x}_{k,t}|\mathbf{x}_{k,t-1}) \prod_{t \in \mathcal{T}_{k,obs}} P(\mathbf{y}_{k,t}|\mathbf{x}_{k,t}) d\mathbf{x}_{1,0} \dots d\mathbf{x}_{K,T_K} \\ - \sum_{i=1}^p \sum_{j=1}^p \lambda_i^{(s)} |A_{i,j}| - \sum_{i=1}^p \sum_{j=1}^m \lambda_i^{(s)} |G_{i,j}| - \sum_{i'=1}^q \sum_{j=1}^p \lambda_{i'}^{(o)} |H_{i',j}|, \end{aligned} \tag{4}$$

where $\lambda_i^{(s)}$ and $\lambda_{i'}^{(o)}$ are the L1 regularization terms for the i th row of system and observation models, respectively.

In the EM algorithm, the conditional expectation of the joint log-likelihood of the complete data set

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}_l) = E[\log P(Y, X|\boldsymbol{\theta})|Y; \boldsymbol{\theta}_l], \tag{5}$$

is iteratively maximized with respect to $\boldsymbol{\theta}$ until convergence, where $\boldsymbol{\theta}_l$ is the parameter vector obtained at the l th (previous) iteration. More details of the proposed algorithm are described in the supplemental materials.

Because the EM-algorithm tries to search for the parameter values that can achieve the local minimum likelihood, we set several initial values and the dimension of the hidden variables p to obtain better models. To select the most plausible model among the estimated results, we apply Bayesian information criterion [17] described as

$$\text{BIC} = -2\log L(Y|\boldsymbol{\theta}) + \text{df}(\boldsymbol{\theta})\log v, \tag{6}$$

$$L(Y|\boldsymbol{\theta}) = \int \prod_{k=1}^K P(\mathbf{x}_{k,0}) \prod_{t \in \mathcal{T}_k} P(\mathbf{x}_{k,t}|\mathbf{x}_{k,t-1}) \prod_{t \in \mathcal{T}_{k,obs}} P(\mathbf{y}_{k,t}|\mathbf{x}_{k,t}) d\mathbf{x}_{1,0} \dots d\mathbf{x}_{K,T_K}, \tag{7}$$

where $df(\theta)$ is the degree of freedom, *i.e.*, the number of non-zero parameters values, and v is the number of observed points. The source code is available at <https://github.com/hase62>.

Results

EHR in Hirosaki cohort study

We apply the method to the data obtained from the Hirosaki Center-of-Innovation (COI) health promotion project, which is a cohort study in Hirosaki City (population is about 10,000) in northern Japan. Every year, about 1,000 people have been checking their health status with free participation. Since the final goal is to extend the life span, it has measured comprehensive items (total about 2,000), such as invasive/non-invasive clinical test values, body composition values, lifestyles, medication, exercise capacity, and cognitive ability. The data actually include many missing observations because individuals did not necessarily participate all check-ups. Thus, for example, if the k th participant visited 2009, 2010, 2013, and 2014's check-ups, we have $Y_k = \{y_{k,1}, y_{k,2}, y_{k,5}, y_{k,6}\}$, where $y_{k,1}$ is the observation data at the first year (2009) of visit. Note that each $y_{k,t}$ ($t = 1, 2, 5, 6$) includes all features. All participants gave written informed consent and the study was approved by IRB in Hirosaki University (Num. 2017-026).

We used the data obtained from 2007 to 2015. Especially in this study, we firstly focused on 1,196 elderly persons aged from 55 to 75 because young participants may have normal blood test values regardless of their lifestyle and we are interested in the association of the effect of lifestyle changes and blood test values in elderly people. We considered exclusion criteria that removed 533 participants who used any medicine because we are interested in the effect of lifestyle changes to the blood test values, but not the effect of medicine. We thought that the blood test values were potentially affected by the medicine. We further excluded 104 participants who participated only one medical check-up between 2007 and 2015 because we cannot obtain any time-series blood test values from them. Finally, the numbers of participants for male and female were 156 and 273, respectively, and thus 429 (35.9%) participants were remained from 1,196 elderly participants.

For baseline information comparison, we prepared histograms of ages and target blood test values of (i) included participants, (ii) excluded participants who used any medicine, and (iii) excluded participants who participated only one medical check-up in the supplementary materials. We then evaluated the differences between (i) and (ii) datasets and between (i) and (iii) datasets by the Wilcoxon rank-sum test. In the former comparison, we hypothesized that some differences should be observed, because dataset (ii) were collected sick participants and they were not our analysis target. From the Wilcoxon rank-sum test between (i) and (ii), we confirmed the difference is significant as considered. Therefore, the participants in (ii) could be affected by using medicine for the treatment and this exclusion criterion seems to be reasonable. On the other hand, we confirmed that the comparison between (i) and (iii) indicated no significant difference (the adjusted threshold of the p -value is $0.05/(39 \times 2) = 6.41E-4$ by the Bonferroni method). Thus, no serious sample bias has been detected.

We focus on 38 blood test values in the collected check-up results and they are handled as $y_{k,t}$ and listed in [Table 1](#). These check-up items are generally collected for the screening of chronic diseases such as Diabetes and high blood pressure. As environmental factors, we prepare body composition values [Bc], lifestyles [Lf] and social status [Ss], and they are handled as $z_{k,t}$ and listed in [Table 2](#). Here, we used Body weight, Abdominal Circumference, W/H-ratio, BMI, and BFP as representative of body composition because they are usually used as easy indicators of obesity, exercise, diet and so on.

In order to show the overall picture of the correlation among blood test values and lifestyles, the correlations of $y_{k,t}$ and $z_{k,t}$ for Male dataset are illustrated in [Fig 1](#). Here, in order

Table 1. The list of observed blood test values in $y_{k,t}$.

Num.	Name	Num.	Name
y1	Systolic Blood Pressure	y20	ALT (GPT)
y2	Diastolic Blood Pressure	y21	Total Protein
y3	PWV Left-Right Average	y22	ALB (Improved BCP)
y4	ABI Left-Right Average	y23	Creatinine
y5	Bone Density (Acoustic Value)	y24	Urea Nitrogen
y6	Bone Density (Z score)	y25	Uric Acid
y7	Bone Density (T score)	y26	Total Cholesterol
y8	Serum Glucose	y27	Triglyceride (TG)
y9	HbA1c (NGSP)	y28	HDL Cholesterol
y10	Muscle Amount (Height Corrected)	y29	LDL Cholesterol
y11	White Blood Cell Count (WBC)	y30	Sodium
y12	Red Blood Cell Count (RBC)	y31	Potassium
y13	Hemoglobin	y32	Chlorine
y14	Hematocrit	y33	Calcium
y15	MCV	y34	Inorganic Phosphorus
y16	MCH	y35	Serum Iron
y17	MCHC	y36	C3
y18	Total Bilirubin	y37	C4
y19	AST (GOT)	y38	nonHDL Cholesterol

<https://doi.org/10.1371/journal.pone.0230172.t001>

Table 2. List of environmental factors representing body composition values [Bc], lifestyles [Lf], and social status [Ss] handled as $z_{k,t}$.

Num.	Name
z1	[Bc] Body Weight
z2	[Bc] Abdominal Circumference
z3	[Bc] Waist / Hip Ratio
z4	[Bc] Body Mass Index
z5	[Bc] Body Fat Percentage
z6	[Lf] Working Days (Days / Week)
z7	[Lf] Sleep Disorder (0:None, 1:Yes)
z8	[Lf] Current Smoking Habits (0:None, 1:Yes)
z9	[Lf] Current Drinking Habits (0:None, 1:Yes)
z10	[Lf] Sleeping Hours
z11	[Lf] Midday Nap (0:None, 1:Yes)
z12	[Lf] Averaged Exercise Hours (1:None, 2:1 Time / Week, 3:2-3 Times / Week, 4:4-5 Times / Week, 5:Everyday)
z13	[Ss] Age at Survey Date
z14	[Ss] Family Structure (#Family members)
z15	[Ss] Marital Status (0:None, 1:Yes)
z16	[Ss] Farmer (0:None, 1:Yes)
z17	[Ss] Final Educational Background (0:Elementary or Junior High or High School, 1:Junior College or Vocational School, 2:University or College)

<https://doi.org/10.1371/journal.pone.0230172.t002>

MALE from 55 to 75

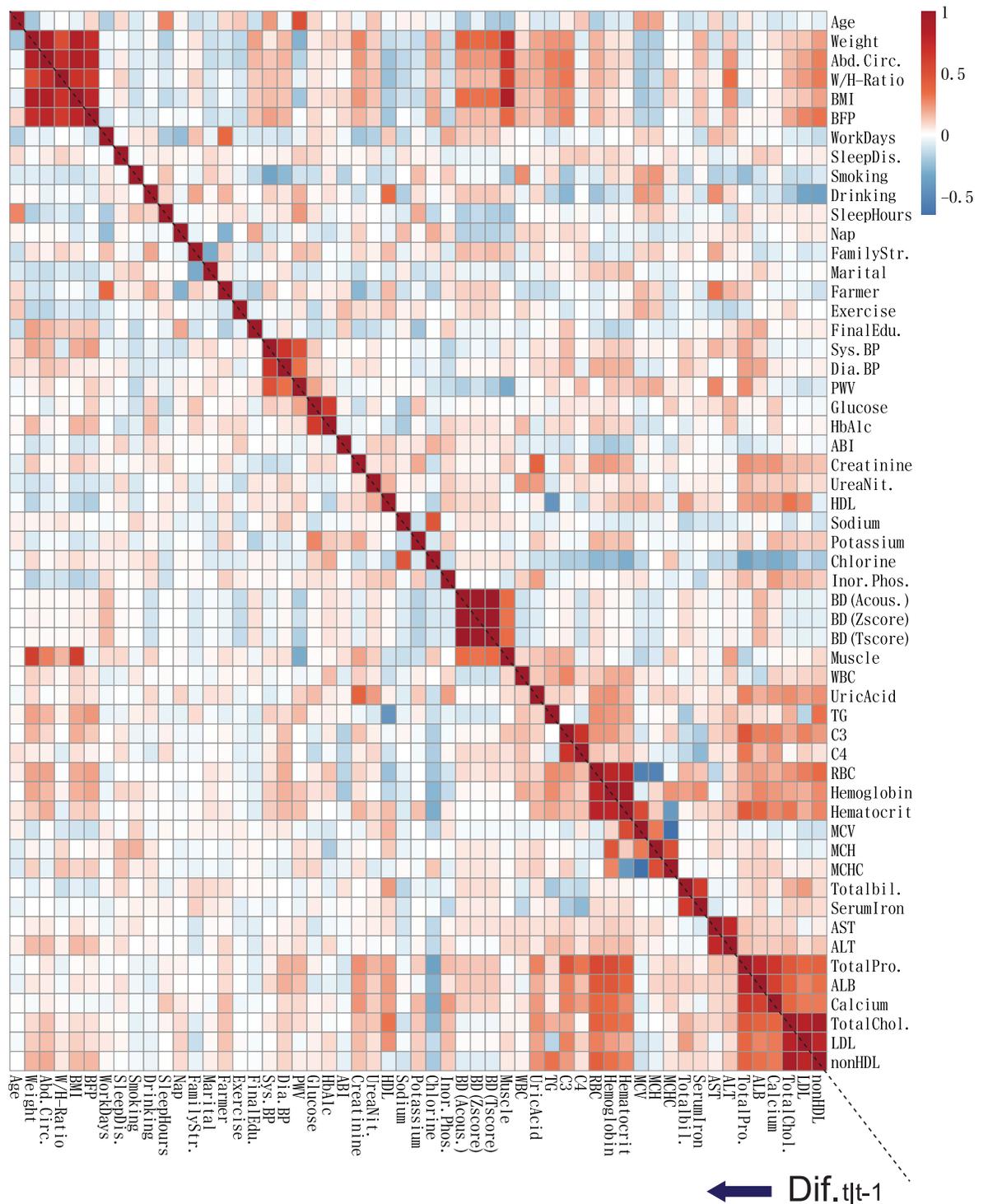


Fig 1. The Spearman correlation coefficients of observed blood test values and environmental factors. The bottom-left part and the upper right part of the heatmap represent the Spearman correlation coefficients among $(y_{k,t} - y_{k,t-1})$ and $(z_{k,t} - z_{k,t-1})$, and among $y_{k,t}$ and $z_{k,t}$ for Male time-series data, respectively.

<https://doi.org/10.1371/journal.pone.0230172.g001>

to illustrate the immediate effect of changes of environmental factors to the changes of blood test values, we illustrated the results of the spearman correlation analysis among $(y_{k,t} - y_{k,t-1})$ and $(z_{k,t} - z_{k,t-1})$, and $y_{k,t}$ and $z_{k,t}$. In the blood test values $y_{k,t}$, we can see that some values are highly correlated, for example, ‘HbA1c and Glucose’, and ‘Total Cholesterols, LDL, and nonHDL’. They should be represented by the same hidden variables in the proposed SSM. Similarly, body composition values are mutually highly correlated, and they are also correlated with Bone Densities (BDs), Muscle Amount, Blood Counts, Cholesterols, and so on. On the other hand, among $(y_{k,t} - y_{k,t-1})$ and $(z_{k,t} - z_{k,t-1})$, they have weaker correlations especially in Waist/Hip ratio (W/H-ratio) than those in $y_{k,t}$ and $z_{k,t}$. Thus, the changes of body composition values and lifestyles cannot immediately affect to blood test values even if some values suddenly change. These correlations enable us to impute missing observation data. For example, high correlations between some observation variables at time t and $t-1$ enable us to predict these variables at time t' , which is a time-point with missing observation data, from the variables at time $t' - 1$. The details of the calculation are written in the supplemental materials.

Inference of internal medical states and their relationship

For the above explained datasets, we applied the proposed SSM to classify the relationships among blood test values and environmental factors. We tested 50 calculations for each system dimension p from 3 to 15 with randomly prepared initial parameter values, and then searched for the value of p that achieves the lowest BIC score as described in Eq (6). Consequently, for both datasets, $p = 14$ was selected as the best dimension. The BIC score is illustrated for each of system dimensions, p , in the supplementary materials.

For each dataset, we next evaluated the effects of $x_{k,t}$ and $z_{k,t}$ to the blood test values $y_{k,t}$. The heatmaps of the estimated sparse matrices H and G are illustrated in Fig 2. This figure shows the effect of $z_{k,t}$ to $x_{k,t}$ as an estimated matrix G in the upper part and that of $x_{k,t}$ to $y_{k,t}$ as an estimated matrix H . Through the results, some clusters exist in the blood tests. We summarize here: ‘Diastolic and Systolic Blood Pressures (BPs), and PWV’ (BP Group), ‘HbA1c and Glucose’ (Diabetes Group), ‘RBC, Hemoglobin, and Hematocrit’ (RBC Group1), ‘MCH and MCV’ (RBC Group2), ‘Total Cholesterols, LDL Chol., HDL Chol, and nonHDL Chol.’ (Dyslipidemia Group), ‘Total Bilirubin and Serum Iron’ (Liver Group1), ‘ALT and AST’ (Liver Group2), and ‘C3, Total Pro., ALB, and Calcium’ (Liver Group 3). These groups are similar to the result of correlation analysis in Fig 1 and can be similarly changed in response to the changes of $x_{k,t}$ and $z_{k,t}$. Also, some hidden variables have similar effect to $y_{k,t}$ in H with partially different components. On the other hand, some blood test values, *e.g.*, ABI and Potassium, are not or weakly regulated by $x_{k,t}$; these blood test values are robust from other blood test values, and the changes of our prepared body composition values and lifestyles. Here, we summarize the characteristics of the results in Table 3.

For each dataset, we then evaluate the effect among $x_{k,t}$ based on the matrix F . Hidden variables that share similar effects in H seems to have feedback structures, but we cannot capture any clear structure. The regulatory relationships are illustrated in the supplemental materials.

Scenario-based blood test value simulation

We compare the predicted and actual time-series blood test values of participants when improving or corrupting their body composition values and lifestyles. In this experiment, we postulate two situations; (i) a healthy person suddenly corrupted their body composition values and lifestyles, and (ii) an unhealthy person started to improve their body composition values and lifestyles. At first, we extracted a healthy male and female, and predict blood test

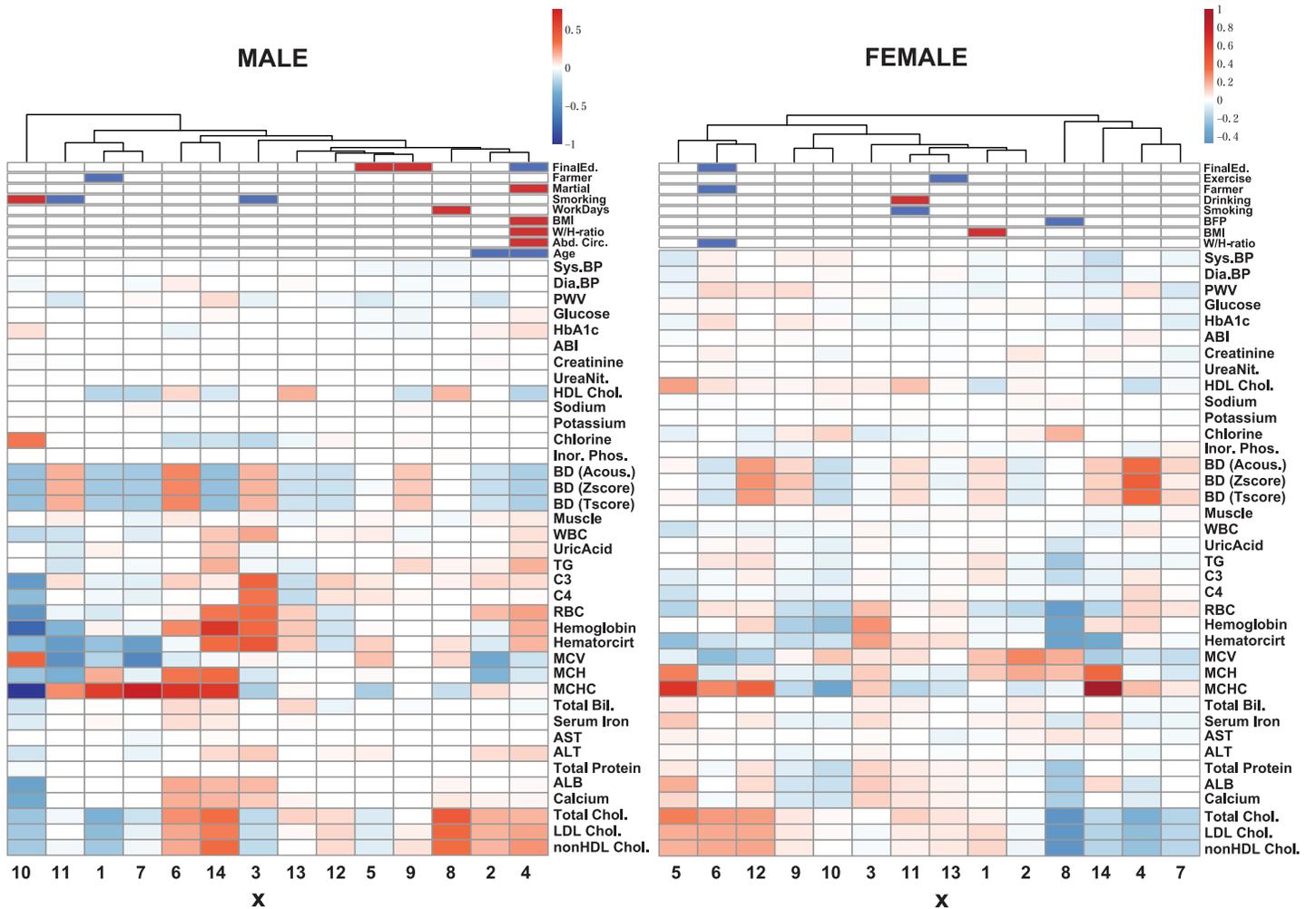


Fig 2. The heatmap of the effect of $z_{k,t}$ to $y_{k,t}$. The heatmap represents the effect of $z_{k,t}$ to $x_{k,t}$ as an estimated matrix G in the upper part and that of $x_{k,t}$ to $y_{k,t}$ as an estimated matrix H in the bottom part, and thus introduces the indirect effect from $z_{k,t}$ to $y_{k,t}$. The numbers on horizontal axis mean the element numbers of the 14-dimensional vector $x_{k,t}$. For example, the column with '3' in the heatmap displays the effect from the 3rd element of the hidden variable $x_{k,t}$ to the observation variables $y_{k,t}$. For H , each value is normalized as $H_{i,j} \times \sqrt{(\sum_{k=1}^K \sum_{t=1}^{T_k} x_{k,t,j}^2)}$ and then $H_{i,j} / \max(|H|)$, where $H_{i,j}$ is the i 'th row and the j 'th column element of H and $\max(|H|)$ is the max absolute value of H . For G , positive and negative elements are illustrated as red and blue, respectively.

<https://doi.org/10.1371/journal.pone.0230172.g002>

values using the estimated parameter values and (i-a) actual $z_{k,t}$ and (i-b) postulated $z_{k,t}$ to conduct them to be unhealthy persons. Here, we consider unhealthy status as Abdominal Circumference (AC) = 95 [cm], W/H-ratio = 1.05, BMI = 30, and Smoking status for male, and W/H-ratio = 0.95, BMI = 30, Body Fat Percentage (BFP) = 35, and Smoking status for female. The simulation results are illustrated in Fig 3.

In these figures, we assume that the body composition values and lifestyles suddenly changed to unhealthy condition at the first medical check-up year. From the simulations, participants in Dyslipidemia Group seem to be the most affected. The male under unhealthy condition gradually increases TG and LDL and finally becomes Dyslipidemia when he is 64 years old. In contrast to the postulated unhealthy condition, actual status at 64 years old is not Dyslipidemic. Furthermore, ALT and Uric Acid are weakly increasing to suspected areas, and HbA1c, BDs, WBC, RBC, Hemoglobin, and Hematocrit are also weakly increasing in the Male

Table 3. Effect of lifestyles for blood test values. The columns (+) and (-) indicate the list of blood test values that are increased and decreased by the corresponding environmental factors in male or female, respectively.

Environmental factors	Male (+)	Male (-)	Female (+)	Female (-)
Farmer	MCV, Hematocrit, Dyslipidemia Group, and BDs	MCH, and MCHC	BDs, MCV, and Hematocrit	Dyslipidemia Group, and MCHC
Increasing Age	BDs, HDL and RBC Group2	Dyslipidemia Group, RBC Group1, and MCHC		
Smoking	Chlorine, and MCV	C3, C4, RBC, and Liver Group 3	MCHC	HDL
Drinking			HDL	MCHC
Higher Education	BDs, and RBC Group2	Dyslipidemia Group, RBC Group1, and MCHC	BDs, MCV, and Hematocrit	Dyslipidemia Group, and MCHC
Married Status	Dyslipidemia Group, and RBC Group1	BDs, HDL, and RBC Group2		
Longer Working Hours	Dyslipidemia Group			
Increasing BMI	Dyslipidemia Group, and RBC Group1	BDs, HDL and RBC Group2	Dyslipidemia Group, and RBC Group2	HDL, and RBC
Increasing W/H-ratio	Dyslipidemia Group, and RBC Group1	BDs, HDL and RBC Group2	BDs, MCV, and Hematocrit	Dyslipidemia Group, and MCHC
Increasing AC	Dyslipidemia Group, and RBC Group1	BDs, HDL and RBC Group2		
Increasing BFP			RBC Group1, TG, Dyslipidemia Group, and Liver Group3	RBC Group2, and Chlorine
Exercise				Dyslipidemia Group

<https://doi.org/10.1371/journal.pone.0230172.t003>

in contrast that they are almost not varied in the Female. In both persons, BPs and HbA1c are not affected by the changes in body composition values and lifestyles used here.

Next, we extracted unhealthy male and female and calculated their predicted blood test values, and (ii-a) actual $z_{k,t}$ and (ii-b) postulated $z_{k,t}$ to conduct them to be healthy persons. Here, we consider healthy conditions as AC = 75[cm], W/H-ratio = 0.80, BMI = 20, and non-smoking status for male and W/H-ratio = 0.70, BMI = 20, BFP = 20, non-smoking, non-drinking and exercise every day. The simulation results are illustrated in Fig 4.

In these experiments, when healthy male participant was changed his lifestyle to unhealthy one; AC, W/H-ratio, and BMI, were changed from less than 75, 0.80, and 20, to 95, 1.05, and 30, respectively. Also, their means in Male participants are approximately 85, 0.9, and 24, respectively. Because each element of $z_{k,t}$ is normalized to mean 0 and variance 1 in the application, the rate of change of the effect by the j th lifestyle ($g_{i,j} \times z_{k,t,j}$) to the i th hidden variable in this case is almost -1, where $g_{i,j}$ is the i th row and the j th column element of G and $z_{k,t,j}$ is the j th element of $z_{k,t}$. On the other hand, smoking and drinking were changed from 0 to 1. The same applies to the opposite case. Similar to the previous experiments, Dyslipidemia Group seems to be the most affected blood test values. In these experiments, the male under unhealthy condition gradually decreases TG and LDL and finally becomes healthy from Dyslipidemia when he is 72 years old. In contrast to the previous experiments, ALT, Uric Acid, HbA1c, BDs, WBC, RBC, Hemoglobin, and Hematocrit are weakly affected both in the Male and Female. In both persons, BPs and HbA1c are not affected.

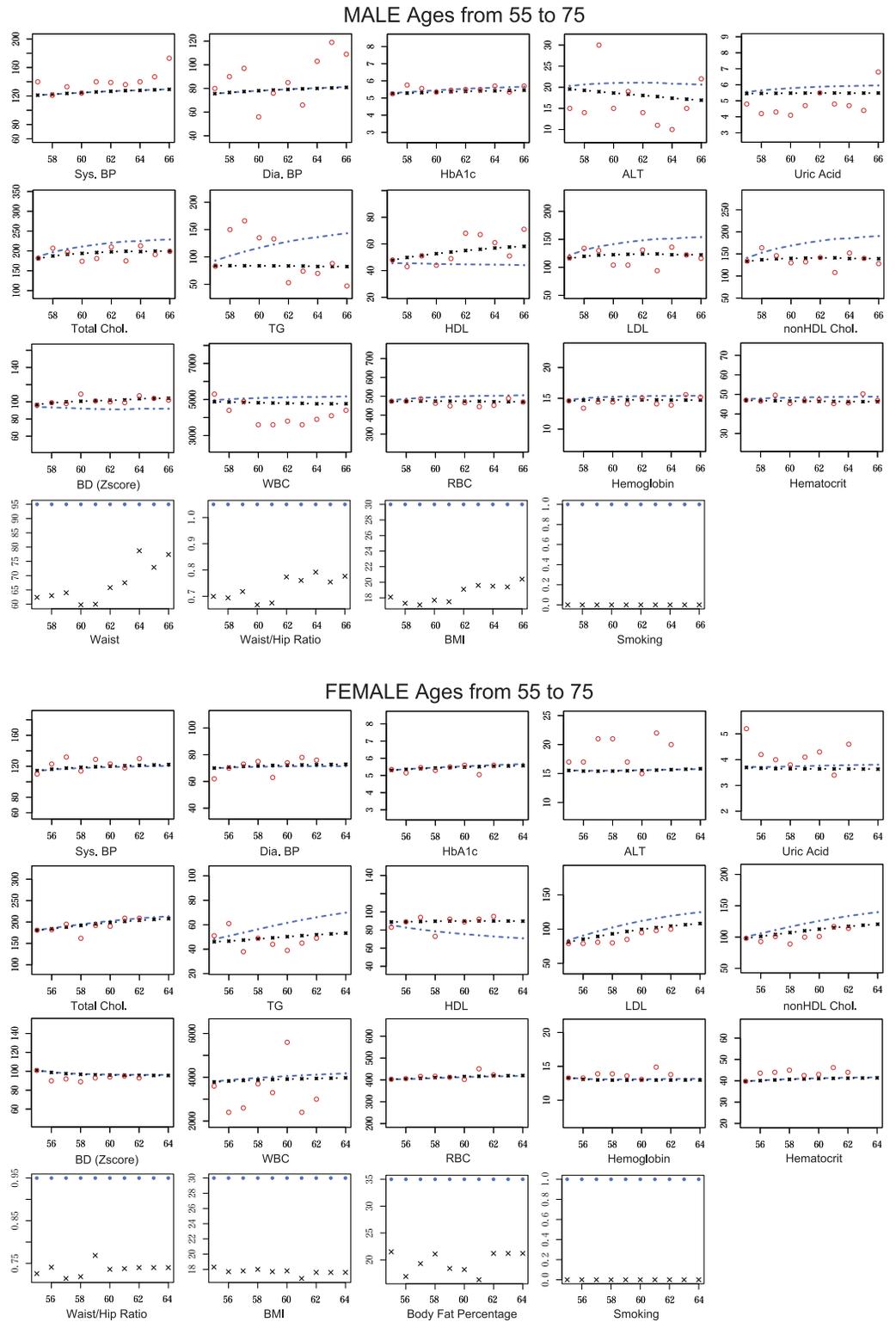


Fig 3. The comparison among predicted and actual blood test values of representative healthy male and female. The predicted values are calculated both under actual and unhealthy conditions for representative healthy male and female. The predicted blood test values are calculated by our estimated state space model and $z_{k,t}$ (actual and unhealthy) without filtering steps. Red circles are observed blood test values, black lines are predicted blood test values under actual condition, and blue lines are predicted blood test values under the unhealthy condition. Black crossed marks are actual

$z_{k,t}$, and blue circles are postulated $z_{k,t}$ under the unhealthy condition. The horizontal and vertical axes indicate the ages of the selected participant at the medical check-up years and the blood test values, respectively. Thus, the youngest age in the horizontal axis means the participant's age at the first year of visit to the medical check-up event. Note that, in the calculation, we consider the observation at the first year of visit as $y_{k,1}$.

<https://doi.org/10.1371/journal.pone.0230172.g003>

Genetic effect analysis

Finally, we analyzed the prediction errors among the predictive and observed values (*observed value—predicted value*). In some participants, there is a clear discrepancy and we assume that they are due to unobserved or unused information such as genetic effects. In the Hiroasaki cohort study, genetic data was also obtained in 2014 and 2015, and we focused on detecting associated SNPs for these discrepancies.

Thus, we firstly created a list of known SNPs from GWAS catalog [18] by indexing names of the used 38 blood tests as queries and obtained 3,924 SNPs, of which 915 SNPs are included in our SNP array. We then (i) performed normalization transformation [19] to the observation values and the prediction errors and (ii) checked the genetic associations of these SNPs with all 38 traits by statistical tests (linear regression analysis under additive, recessive, and dominant models with adjustment of age and BMI). To extract firm associations, we set the adjusted threshold for p value lower than $0.05/915 = 5.46E - 5$ by the Bonferroni method. Note that, in this analysis, some observed values can be generated by the same participants diagnosed at different years and, in these cases, we used mean values to avoid the inflation of p -values. Consequently, we obtained SNPs that were strongly associated to the prediction errors in both cases. Parts of the results are illustrated in Fig 5.

It is known that rs662799 is associated with higher TG and cholesterol levels [20] and participants who have this SNP were predicted to exhibit positive prediction errors for nonHDL cholesterol levels ($p = 5.25E - 5$) in the male result. Additionally, rs887829 is popularly associated with total bilirubin levels [21] and participants who have this SNP were predicted to exhibit positive prediction errors for total bilirubin levels ($p = 4.17E - 8$) in the female result. However, in contrast to the results of rs887829, the strong association of rs662799 with nonHDL cholesterol levels could not be detected when using the observation values ($p = 2.34E - 2$). In the middle and right of this figure, novel effects of SNPs were suggested in the associations of the prediction errors. rs4607103 and rs3817198 are known to have an association with type II Diabetes [22] and Breast cancer [23], respectively, but associations with total protein levels ($p = 3.23E - 5$) and PWV ($p = 3.63E - 5$) were indicated in the male results, respectively. Similarly, rs907612 and rs2371767 were indicated to have associations with Uric Acid levels ($p = 1.00E - 5$) and HDL cholesterol levels ($p = 2.57E - 5$) in the female results, respectively, but these SNPs are known to have associations with monocyte and granulocyte [24] and Waist/Hip-ratio [25], respectively. Among them, strong associations of rs4607103, rs3817198 and rs2371767 could not be detected when using observation values.

Discussion and conclusion

Generally, we could control some blood test values in response to the changes of body composition values, lifestyles, and social status. In this study, we focus on chronic diseases such as Diabetes and high blood pressure and selected 38 blood test values listed in Table 1. Then, we defined $z_{k,t}$ as clinical target values to be controlled as listed in Table 2 and clarified their effect to the blood test values. Body composition values are considered as endpoints of both exercise and eating habits because it is difficult to directly clarify eating habits. Then, we assume that the changes in body composition values could partially reflect the effects of eating habits. Actually, other blood tests, e.g., γ GT and C-reactive protein, could be effective items to investigate

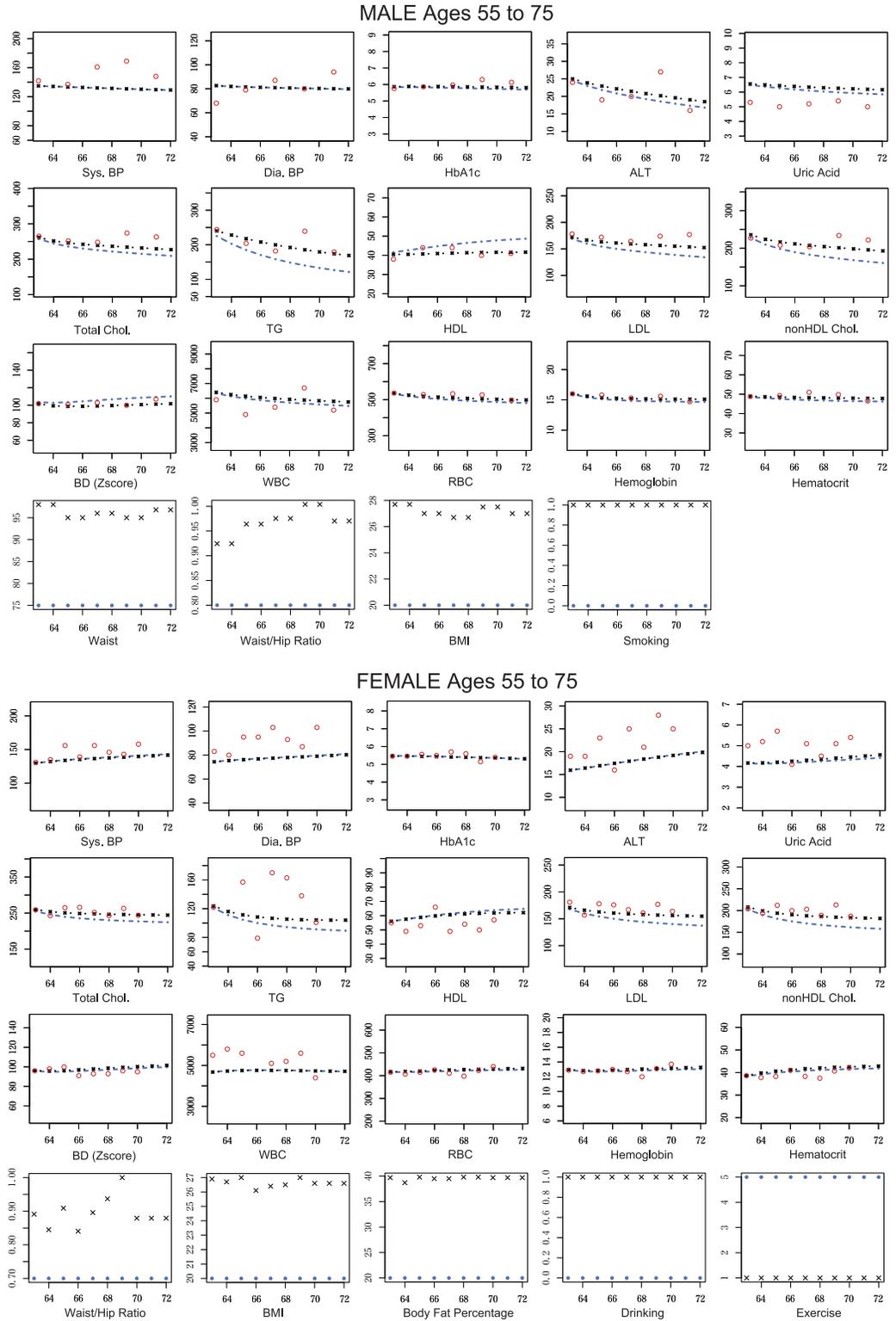


Fig 4. The comparison among predicted and actual blood test values of representative unhealthy male and female. The predicted values are calculated both under actual and healthy conditions for representative unhealthy male and female. The predicted blood test values are calculated by our estimated state space model and $z_{k,t}$ (actual and healthy) without filtering steps. Red circles are observed blood test values, black lines are predicted blood test values under actual condition, and blue lines are predicted blood test values under the unhealthy condition. Black crossed marks are actual

$z_{k,t}$, and blue circles are postulated $z_{k,t}$ under the healthy condition. The horizontal and vertical axes indicate the ages of the selected participant at the medical check-up years and the blood test values, respectively. Thus, the youngest age in the horizontal axis means the participant's age at the first year of visit to the medical check-up event. Note that, in the calculation, we consider the observation at the first year of visit as $y_{k,1}$.

<https://doi.org/10.1371/journal.pone.0230172.g004>

chronic diseases, but they have been collected recently in this cohort. On the other hand, it may be interesting to analyze blood test values such as vitamins and hormones for other diseases.

From Fig 1, we can see that many blood test values are highly correlated among them, as well as correlated to body composition values, lifestyles, and social status. In contrast, focusing on their year-differences $y_{k,t} - y_{k,t-1}$ and $z_{k,t} - z_{k,t-1}$, absolute correlation values are lower than the correlations among $y_{k,t}$ and $z_{k,t}$. This indicates that blood test values are not suddenly changed according to the changes of other values (except for the case where they indicate almost the same vital condition, such as in HbA1c and glucose). Thus, we can assume that lifestyle changes gradually impact blood test values.

From the results of estimated matrices H and G in the proposed SSM, similar to the results obtained in Fig 1, the blood test values are clustered into categories such as RBC and Dyslipidemia Groups. They are regulated by the same hidden variables $x_{k,t}$ and environmental factors $z_{k,t}$ and assumed to be similarly varied. Additionally, we can see that body composition values can control some indicators for chronic diseases, e.g., Dyslipidemia Groups, but simultaneously and oppositely regulate RBC Groups. It is quite natural that chemical compounds are mutually regulated or share the same chemical pathways and they should be simultaneously affected even if we change only a part of the blood test values.

Then, we postulated the case of healthy or unhealthy persons suddenly changing their body composition values and lifestyles and predicted time-transitions of their blood test values. In both males and females, indicators for Dyslipidemia such as total cholesterol, LDL, HDL, and TG, are strongly affected by changes in body composition values such as AC, W/H-ratio, and BMI. These results are intuitively natural because serum cholesterol levels show the lipid profiles in blood whereas body composition values show subcutaneous and visceral fat. Thus, the amount of cholesterol should have a mutual relationship with the body composition values. On the other hand, HbA1c and BPs are slightly affected by the changes of body composition values and lifestyle changes. Exercise could be one of several factors related to and prevent diabetes; however, even if you exercise, you still may get diabetes due to other factors. For example, HbA1c levels can increase if the participants started exercising but were still eating and drinking too much. As a result, their clear relationship cannot be detected in the experiments and we consider these values could be gradually changed over time based on their entire lifestyle. In contrast, ALT, Uric Acid, and BDs values are varied only in males in relation to the changes of indicators for Dyslipidemia. We assume that these blood test values are likely to be changed in males and these values are stable enough in females to reveal the change of trends in our dataset. Consequently, we conclude that the body composition values and lifestyle changes can be a clinical target of Dyslipidemia and the risk of the disease might be controlled by keeping them within regular ranges. Our quantitative results using the mathematical simulation are consistent with types of intuitively reasonable knowledge and capable of updating them in more detail. For other blood tests that are targets of other diseases, such as Diabetes, we have to add other controllable targets in environmental factors. For example, time of eating dinner and composition of the meal can be good factors.

Although the observation data are well predicted for the most part, if predicted values still vary from the observation data in only some participants, they do not share the estimated

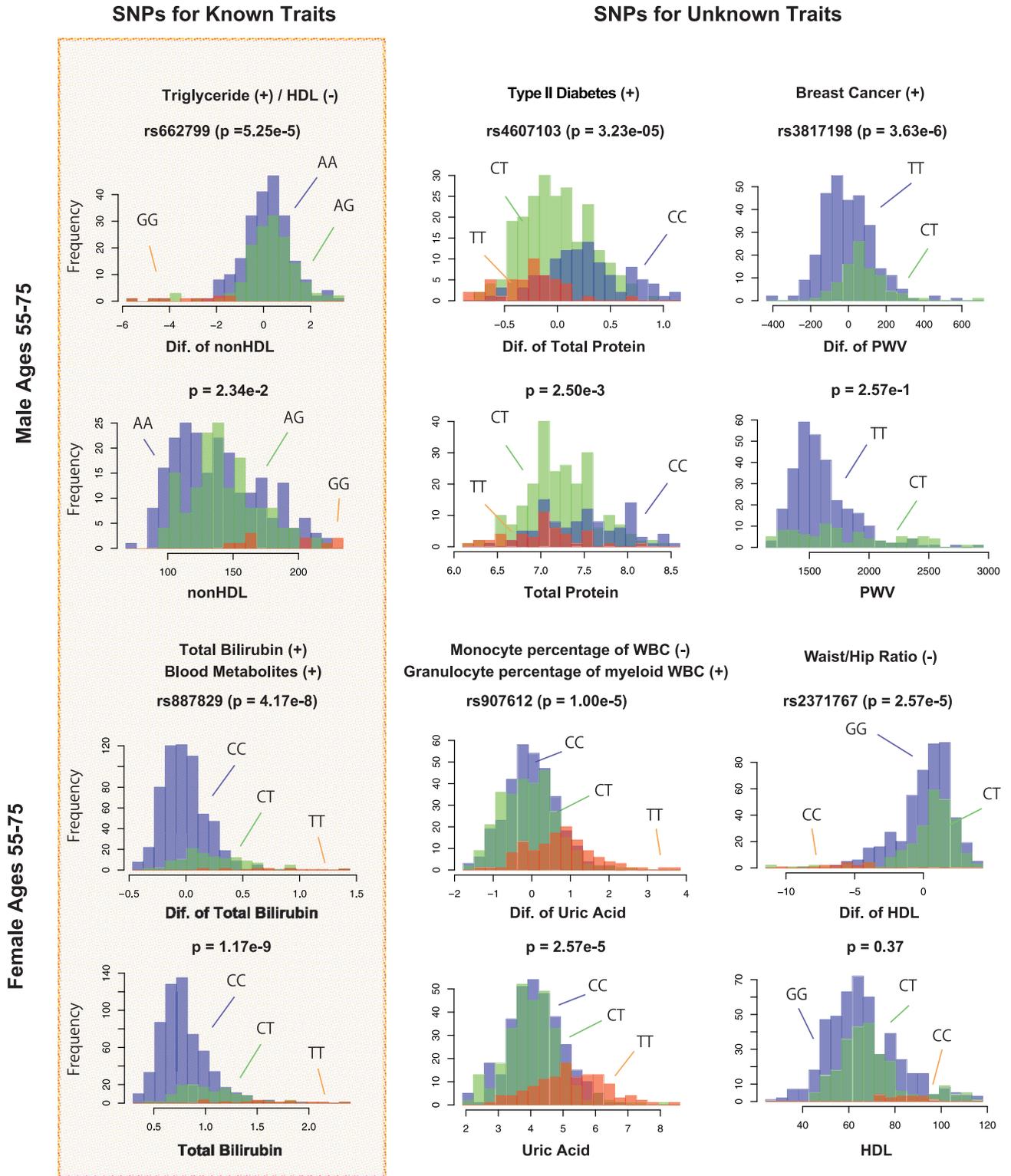


Fig 5. Histograms of the observed blood test values and prediction errors among predicted blood test values of our proposed models and the observed blood test values. The first and the second rows display the results for the Male and the third and the fourth rows display the results for the

Female. The left histograms indicate the association of known SNPs and known traits. The middle and right histograms indicate the association of known SNPs and unknown traits. The sign on the right of the blood tests indicates the positive (+) and negative (-) effects from SNPs. Blue, green, and red histograms show the cases of major homo, hetero, and minor risk alleles, respectively.

<https://doi.org/10.1371/journal.pone.0230172.g005>

relationships represented by A , H , and G . In this case, we consider that other factors, which are not included in the model, could affect to their relationships or observation variables. To clarify these factors, finally, we evaluated genetic effects for the correction of prediction errors among the predictive values under the participant's actual conditions and observed values. We discovered some interesting findings about detected SNPs. In the results of SNPs with known traits, rs662799 has a strong association with higher TG and cholesterol levels [20] and this association could be reproduced when using the prediction errors ($p = 5.25E - 5$). However, a strong association could not be detected when using observation values ($p = 2.34E - 2$). In the results of SNPs with novel traits, the association of rs907612 with uric acid levels was detected by using either prediction errors or observation values. However, associations of rs4607103, rs3817198 and rs2371767 with total protein, PWV and HDL cholesterol, respectively, were detected only by using prediction errors. rs4607103 was suggested to be associated with type II Diabetes [22] and Diabetes is also known to be associated with the increase of total protein. Similarly, rs3817198 had an association with breast cancer [23] and recent studies indicated the association of breast cancer and PWV [26, 27]. The association of rs2371767 to waist/hip-ratio was previously indicated [25] and we can assume that the waist/hip-ratio has an association with the levels of HDL cholesterol because both values can be used as indicators for metabolic syndrome. Thus, these associations seem to be novel findings of genetic effects. Moreover, similar to our experiments, rs907612 and rs2371767 were recently obtained in previous studies [24, 25] by conditioning several covariates, *e.g.*, age, exercise, smoking history and drinking history, using a large number of participants (both studies utilized approximately 200,000 participants). Such approaches could clarify the minute effects of SNPs that could not be detected using traditional genome wide association studies, which condition a few covariates or principal components and capture SNPs with higher effect size. Because our approach can utilize time-series information for detecting associated SNPs, it could especially enhance the detection power in contrast to the traditional non-time series approaches.

The extension of the model to include the presence of SNPs enables us to evaluate their effect to the blood test values and to predict them more accurately. However, such extension makes the calculation heavier depending on the dimension of $z_{k,t}$ and it is difficult to include all obtained SNPs in the proposed model. Thus, extended models or more efficient algorithm might be required to include SNPs effects. For example, the calculation using linear mixed models are computationally efficient and it can be applied with designing appropriate covariance structure for time-series data. In addition, since random effects could correct for population structure and family relatedness also in SSM, it can be extensions of the proposed model. Some other factors can also affect to the changes of blood test values. The analysis including SNP effects, random effects, and other factors is one of future works. Our proposed model can express linear relationships under the Gaussian assumption among blood test values and environmental factors, but many natural processes, *e.g.*, chemical reaction networks and biological signal processing, are generally represented as nonlinear structures. Moreover, the observation noises of some blood test values could not depend on the Gaussian noises, for example, TG levels can suddenly increase due to eating habits before medical check-up. The use of a nonlinear structure and non-Gaussian assumption has the potential to clarify more detailed relationship among them.

Supporting information

S1 Appendix. A solution to infer parameter values of the proposed model. The detailed solution of estimating parameter values using the EM-algorithm for the proposed model. (PDF)

S1 Fig. Histogram of ages and blood test values of included and excluded participants. The histograms of ages and 38 observation values $y_{k,t}$ of included and excluded samples for both dataset. (PDF)

S2 Fig. The lowest BIC score of each system dimension p . The lowest BIC score of each system dimension p ($p = 1, \dots, 15$) for both dataset. (PDF)

S3 Fig. A graphical view of regulatory matrix F . The regulatory relationships among $x_{k,t}$ and $z_{k,t}$ generated from F for both dataset. (PDF)

Author Contributions

Conceptualization: Takanori Hasegawa.

Data curation: Kaori Sawada, Kenichi Kawatani, Koichi Murashita.

Formal analysis: Kaori Sawada, Kenichi Kawatani, Koichi Murashita.

Funding acquisition: Shigeyuki Nakaji, Seiya Imoto.

Investigation: Takanori Hasegawa, Masanori Kakuta.

Methodology: Takanori Hasegawa, Rui Yamaguchi, Masanori Kakuta, Seiya Imoto.

Project administration: Shigeyuki Nakaji.

Supervision: Shigeyuki Nakaji, Seiya Imoto.

Writing – original draft: Takanori Hasegawa.

Writing – review & editing: Rui Yamaguchi, Masanori Kakuta, Kaori Sawada, Kenichi Kawatani, Koichi Murashita, Shigeyuki Nakaji, Seiya Imoto.

References

1. Melnyk I, Yadav P, Steinbach M, Srivastava J, Kumar V, Banerjee A. Detection of Precursors to Aviation Safety Incidents Due to Human Factors. In: 2013 IEEE 13th International Conference on Data Mining Workshops; 2013. p. 407–412.
2. Rana S, Gupta S, Phung D, Venkatesh S. A predictive framework for modeling healthcare data with evolving clinical interventions. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 2015; 8(3):162–182. <https://doi.org/10.1002/sam.11262>
3. Barenco M, Tomescu D, Brewer D, Callard R, Stark J, Hubank M. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*. 2006; 7(3):R25+. <https://doi.org/10.1186/gb-2006-7-3-r25> PMID: 16584535
4. Beal MJ, Falciani F, Ghahramani Z, Rangel C, Wild DL. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*. 2005; 21:349–356. <https://doi.org/10.1093/bioinformatics/bti014> PMID: 15353451
5. Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotharan E, Gaiba A, et al. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*. 2004; 20:1361–1372. <https://doi.org/10.1093/bioinformatics/bth093> PMID: 14962938

6. Nachimuthu SK, Wong A, Haug PJ. Modeling Glucose Homeostasis and Insulin Dosing in an Intensive Care Unit using Dynamic Bayesian Networks. *AMIA Annual Symposium Proceedings*. 2010; 2010:532–536. PMID: [21347035](https://pubmed.ncbi.nlm.nih.gov/21347035/)
7. Sandri M, Berchiolla P, Baldi I, Gregori D, Blasi RAD. Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. *Journal of Biomedical Informatics*. 2014; 48:106–113. <https://doi.org/10.1016/j.jbi.2013.12.008> PMID: [24361388](https://pubmed.ncbi.nlm.nih.gov/24361388/)
8. Gatti E, Luciani D, Stella F. A continuous time Bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*. 2012; 24(4):496–515. <https://doi.org/10.1007/s10696-011-9131-2>
9. Peelen L, de Keizer NF, de Jonge E, Bosman RJ, Abu-Hanna A, Peek N. Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit. *Journal of Biomedical Informatics*. 2010; 43(2):273–286. <https://doi.org/10.1016/j.jbi.2009.10.002> PMID: [19874913](https://pubmed.ncbi.nlm.nih.gov/19874913/)
10. Ding X, Chen W, Wassell I. Sparsity-fused Kalman filtering for reconstruction of dynamic sparse signals. In: 2015 IEEE International Conference on Communications (ICC); 2015. p. 6675–6680.
11. Dong CY, Shin D, Joo S, Nam Y, Cho KH. Identification of feedback loops in neural networks based on multi-step Granger causality. *Bioinformatics*. 2012; 28(16):2146–2153. <https://doi.org/10.1093/bioinformatics/bts354> PMID: [22730429](https://pubmed.ncbi.nlm.nih.gov/22730429/)
12. Hasegawa T, Yamaguchi R, Nagasaki M, Miyano S, Imoto S. Inference of Gene Regulatory Networks Incorporating Multi-Source Biological Knowledge via a State Space Model with ℓ_1 Regularization. *PLoS ONE*. 2014; 9(8):e105942. <https://doi.org/10.1371/journal.pone.0105942> PMID: [25162401](https://pubmed.ncbi.nlm.nih.gov/25162401/)
13. Asif HMS, Sanguinetti G. Large-scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics*. 2011; 27(9):1277–1283. <https://doi.org/10.1093/bioinformatics/btr113> PMID: [21367870](https://pubmed.ncbi.nlm.nih.gov/21367870/)
14. Lillacci G, Khammash M. Parameter Estimation and Model Selection in Computational Biology. *PLoS Computational Biology*. 2010; 6(3):e1000696. <https://doi.org/10.1371/journal.pcbi.1000696> PMID: [20221262](https://pubmed.ncbi.nlm.nih.gov/20221262/)
15. Hirose O, Yoshida R, Imoto S, Yamaguchi R, Higuchi T, Charnock-Jones DS, et al. Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models. *Bioinformatics*. 2008; 24:932–942. <https://doi.org/10.1093/bioinformatics/btm639> PMID: [18292116](https://pubmed.ncbi.nlm.nih.gov/18292116/)
16. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*. 1977; 39:1–38.
17. Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. *The Annals of Statistics*. 2007; 35(5):2173–2192. <https://doi.org/10.1214/009053607000000127>
18. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*. 2017; 45(D1):D896–D901. <https://doi.org/10.1093/nar/gkw1133> PMID: [27899670](https://pubmed.ncbi.nlm.nih.gov/27899670/)
19. Peterson RA, Cavanaugh JE. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*. 2019; p. 1–16. <https://doi.org/10.1080/02664763.2019.1630372>
20. Wu Y, Marvelle AF, Li J, Croteau-Chonka DC, Feranil AB, Kuzawa CW, et al. Genetic association with lipids in Filipinos: waist circumference modifies an APOA5 effect on triglyceride levels. *Journal of Lipid Research*. 2013; 54(11):3198–3205. <https://doi.org/10.1194/jlr.P042077> PMID: [24023260](https://pubmed.ncbi.nlm.nih.gov/24023260/)
21. Milton JN, Sebastiani P, Solovieff N, Hartley SW, Bhatnagar P, Arking DE, et al. A Genome-Wide Association Study of Total Bilirubin and Cholelithiasis Risk in Sickle Cell Anemia. *PLOS ONE*. 2012; 7(4):1–10. <https://doi.org/10.1371/journal.pone.0034741>
22. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*. 2008; 40:638 EP –. <https://doi.org/10.1038/ng.120> PMID: [18372903](https://pubmed.ncbi.nlm.nih.gov/18372903/)
23. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007; 447:1087 EP –. <https://doi.org/10.1038/nature05887> PMID: [17529967](https://pubmed.ncbi.nlm.nih.gov/17529967/)
24. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016; 167(5):1415–1429.e19. <https://doi.org/10.1016/j.cell.2016.10.042> PMID: [27863252](https://pubmed.ncbi.nlm.nih.gov/27863252/)
25. Graff M, Scott RA, Justice AE, Young KL, Feitosa MF, Barata L, et al. Genome-wide physical activity interactions in adiposity—A meta-analysis of 200,452 adults. *PLOS Genetics*. 2017; 13(4):1–26. <https://doi.org/10.1371/journal.pgen.1006528>

26. Grover S, Lou PW, Bradbrook C, Cheong K, Kotasek D, Leong DP, et al. Early and late changes in markers of aortic stiffness with breast cancer therapy. *Internal Medicine Journal*. 2015; 45(2):140–147. <https://doi.org/10.1111/imj.12645> PMID: 25404097
27. Yersal Ozlem, A H, M N, Eryilmaz Ufuk, Barutca S. Arterial Stiffness in Breast Cancer Patients Treated with Anthracycline and Trastuzumab-Based Regimens. *Cardiology Research and Practice*. 2018; 2018 (5352914):1–6. <https://doi.org/10.1155/2018/5352914>