Hypertension Research
https://doi.org/10.1038/s41440-026-02715-4

ARTICLE

An interpretable combinatorial data-mining framework for predicting new-onset hypertension in the general population

Yohei Miyashita[1] · Naoki Kimoto[2,3] · Kohsuke Onoue[3] · Yutaka Yata[4] · Takeshi Aketa[5] · Masami Yabumoto[4] · Takashi Washio[6] · Seiji Takashima[2,7] · Masafumi Kitakaze[1,4,7]

Received: 16 March 2026 / Revised: 18 May 2026 / Accepted: 10 June 2026

Abstract

We previously established an interpretable combinatorial data-mining framework to identify combinations of clinical factors predictive of heart failure. Because hypertension (HT) is a major contributor to heart failure, accurate prediction of new-onset HT is critically important for prevention. To identify combinations of clinical factors predictive of HT onset using a novel limitless-arity multiple-testing procedure (LAMP) and to estimate the probability of developing HT. We analyzed 2,610,286 individuals without HT who underwent annual health check-ups starting in 2005–2015 and were followed for 5 consecutive years without missing data. Using the LAMP method, we systematically identified statistically significant combinations of fewer than four clinical factors associated with HT onset. Among 28,618 subjects used for rule discovery, 4802 combinations predictive of HT onset were identified. The remaining 2,581,668 individuals were classified into one group with no predictive combinations (G0) and 20 groups (G1–G20) according to increasing numbers of predictive combinations. The incidence of HT increased stepwise with the number of predictive combinations, as confirmed by Kaplan–Meier analyses ($p < 0.001$). Receiver-operating characteristic analysis demonstrated a moderate discriminative performance (area under the curve = 0.69). We identified combinations of routine clinical parameters that predict new-onset HT in the general population. A greater number of matching predictive combinations was associated with a proportionally higher probability of developing HT. This interpretable combinatorial data-mining framework may enable risk stratification for HT and support early preventive strategies.

Keywords Artificial intelligence · Forecasting · Digital hypertension · Prediction algorithms · Risk assessment

Introduction

Hypertension (HT) imposes a substantial burden on the heart and vasculature, leading to cardiovascular disease and ultimately progressing to heart failure (HF) [1–3]. Therefore, the ability to predict the onset of HT and to identify habitual and lifestyle-related factors that provoke HT among individuals without established HT is critically important for preventing its development. Moreover, identifying previously unrecognized or seemingly unrelated factors that contribute to HT onset may further enhance preventive strategies, although early and strict blood pressure control remains the cornerstone for reducing cardiovascular risk [4, 5].

These authors contributed equally: Yohei Miyashita, Naoki Kimoto

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41440-026-02715-4.

✉ Masafumi Kitakaze
kitakaze@zf6.so-net.ne.jp

[1] Department of Cardiovascular Medicine, Osaka University Graduate School of Medicine, 2-2 Yamadaoka, Suita, Osaka, Japan

[2] Department of Medical Biochemistry, Osaka University Graduate School of Medicine/Frontier Biosciences, 2-2, Yamadaoka, Suita, Osaka 565-0871, Japan

[3] Non Profit Organization Think of Medicine in Science, 3-7-11, Minamisumiyoshi, Sumiyoshi-ku, Osaka, Japan

[4] Hanwa Memorial Hospital, 3-5-8 Minamisumiyoshi, Sumiyoshi-ku, Osaka, Japan

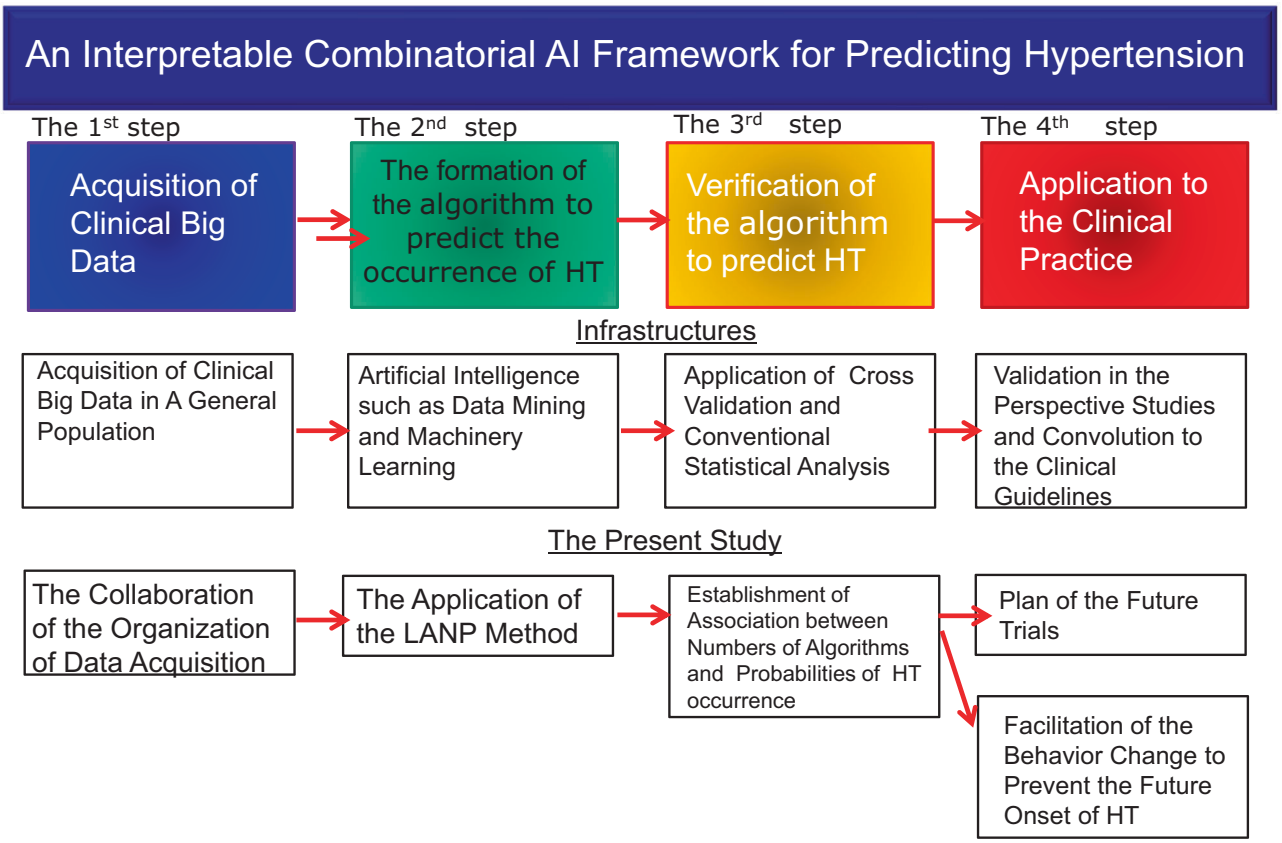
[5] ASCLEPIUS Inc. 3-6-2 Minamisumiyoshi, Sumiyoshi-ku, Osaka, Japan

[6] The Institute of Scientific and Industrial Research, Osaka University, 1-1 Yamadaoka, Suita, Osaka, Japan

[7] The Osaka Medical Research Foundation for Intractable Diseases, 2-6-29 Abikohigashi, Sumiyoshi-ku, Osaka, Japan

Published online: 22 June 2026

Graphical Abstract



Lifestyle-related conditions such as obesity, physical inactivity, excessive alcohol consumption, and high salt intake are well established contributors to HT [6, 7]. However, accurately predicting HT onset in individual subjects remains challenging, because HT arises from the complex interplay of multiple clinical, medical, physical, and habitual factors, rather than from a single dominant cause, as seen in monogenic diseases [8]. This multifactorial complexity highlights the need for a quantitative framework capable of predicting HT onset and identifying high-risk individuals within the general population. Although machine learning–based models have shown promising performance in predicting incident HT, a systematic review has noted substantial methodological heterogeneity and limited explainability across existing studies [9]. Recently, we implemented a novel statistical approach that enables exhaustive analysis of all statistically significant combinations of clinical parameters using a limitless-arity multiple-testing procedure (LAMP) [10–12]. Applying this framework, we successfully identified combinations of clinical factors associated with worsening HF severity in patients with HF and with increased risk of HF onset in the general population [11]. These findings suggest that a combinatorial approach may be particularly well suited for elucidating complex disease phenotypes such as HT.

To translate this framework to the prediction of HT onset, we first recruited Japanese individuals who underwent annual, consecutive health check-ups over a 5-year period, enabling the identification of subjects who did or did not develop HT. This study population was used to explore combinations of factors contributing to HT onset and to examine whether such combinations could predict HT development over 5 years. Second, using ~30,000 individuals from the general population, we identified predictive combinations of clinical, medical, physical, and habitual parameters associated with HT onset using the LAMP method. Finally, we evaluated whether an increasing number of matched predictive combinations was associated with a higher probability of HT occurrence in ~2.6 million individuals from the general population. Through these sequential analyses, we sought to establish an accurate and quantitative framework for predicting HT onset and to clarify the combinatorial factors underlying its pathophysiology in the general population.

Methods

Study design and participants

This retrospective observational study was conducted in Japan using healthcare insurance claims data obtained from the Japan Medical Data Center (JMDC) Inc. (Tokyo, Japan). The JMDC database contains standardized eligibility and claims data provided by multiple health insurance societies and includes ~2.6 million insured individuals, comprising employees of general corporations and their family members. The database captures all medical treatments received by insured individuals across all healthcare facilities, providing a comprehensive longitudinal record of medical care.

All data were anonymized using unlinkable anonymization after removal of decoding indices. The study protocol was approved by the ethics committee of Hanwa Memorial Hospital (approval number: M2024-2). In accordance with the Japanese Ethical Guidelines for Clinical Research, informed consent was waived because of the retrospective observational design. Instead, JMDC Inc. issued a public disclosure as required by the ethics committee and national guidelines. This study was conducted in accordance with the Declaration of Helsinki and the Japanese Ethical Guidelines for Clinical Research.

Protocol 1: identification of predictive combinations using LAMP

We randomly selected 28,618 Japanese individuals without a diagnosis of HT at database entry during 2005–2015 who had complete data for five consecutive years of follow-up. From this cohort, 289 clinical, medical, physical, and habitual parameters assessed at baseline were analyzed. These included sex; age at entry; urinary glucose and protein levels (borderline, 1 +, 2 +, 3 +, or 4 +); plasma LDL cholesterol, HDL cholesterol, and triglyceride levels (mg/L); hemoglobin A1c (%); body mass index; systolic and diastolic blood pressure (mmHg); plasma uric acid (mg/L); fasting plasma glucose (mg/dL); plasma alanine aminotransferase, aspartate aminotransferase, and γ-glutamyl transpeptidase levels (IU/L); abdominal circumference (cm); red blood cell count ($\times 10^4/\mu L$) and hemoglobin concentration (g/dL); chest radiograph findings (A: normal; B: minor changes without need for follow-up; C: follow-up required; H: treatment required); electrocardiographic findings (A: normal; B: minor changes without need for follow-up; C: follow-up required; H: treatment required); visual display terminal work status (A: follow-up required; B: minor changes without need for follow-up; C: normal); lifestyle interview items (current smoking status; ≥30 minutes of daily exercise; body weight change >2 kg within 1 year; alcohol consumption categorized as daily, occasional, or none); and prescription records. Among prescription details included among the baseline clinical parameters, the variable of "use of antihypertensive medication: No" reflects baseline treatment status prior to a formal diagnosis of HT. Of course, all participants were free of diagnosed HT at baseline; however, this variable does not preclude the presence of borderline or high-normal blood pressure values within the non-hypertensive range.

All data underwent rigorous quality control and cleaning procedures. The occurrence of HT was monitored until 2020. HT was diagnosed by HT specialists and general practitioners according to the Japanese diagnostic criteria [13, 14]. Subjects were categorized according to the presence or absence of HT onset during the five-year follow-up period (Table 1).

To identify statistically significant single factors or combinations of factors associated with HT onset, we applied the LAMP.(12) Each subject was represented by a vector of clinical factors and a binary class label indicating HT occurrence. These data were organized into a data table D consisting of N rows and M factors per row. LAMP employs Fisher's exact test to evaluate hypotheses defined by combinations of factors and their association with the outcome class.

Here, the hypothesis was based on a combination of class labels and conditions defined as a subset of the M factors in D. As the condition of the uncovered significant hypothesis may include any number of factors from 1 to M, the term "limitless-arity" has been used to describe this method. Accordingly, LAMP applies a highly efficient search algorithm to quickly and completely derive significant hypotheses from 2^M candidates.

If k is the number of all hypotheses for which the conditions exceed or remain equal to σ objects in D ($\sigma < N$), the relationship between k and σ ($k = k_D(\sigma)$) depends on D but is always antimonotonic because fewer hypothesis conditions remain true at a higher frequency of D. Although the formula of $k_D(\sigma)$ is not analytically determined, LAMP includes a mining algorithm to efficiently derive all k hypothesis conditions under a given σ. Bonferroni correction sets a boundary for the familywise error rate of the false negatives in multiple tests at less than 1 significance level α by correcting the level to $\alpha/k_D(\sigma)$. Bonferroni correction can be used as a standard multiple-testing procedure for the k hypotheses. Note that this level is monotonic to σ as $k_D(\sigma)$ is antimonotonic. If we use a very small set value for σ for a complete search of the significant hypotheses, $\alpha/k_D(\sigma)$ is extremely small because $k_D(\sigma)$ approaches 2^M. In this scenario, almost no hypotheses will be accepted as significant. Conversely, if the set values of σ and, consequently, $\alpha/k_D(\sigma)$ are too large, $k_D(\sigma)$ will be very small and some significant hypothesis conditions will be missed. To

Table 1 Clinical characteristics of test and validation cohorts at baseline with or without the occurrence of hypertension

Variables	The cohort for LAMP analyses			The validation cohort	Statisitical significance between each group		
	Total ($N = 28,618$)	Without hypertension ($N = 27,773$)	With hypertension ($N = 845$)	($N = 2,581,668$)	Total cohort vs. cohort without hypertension	Total cohort vs. cohort with hypertension	Cohorts with hypertension vs. without hypertension
Women, *n* (%)	12,433 (43.4%)	12,176 (43.7%)	295 (34.9%)	1,039,544 (40.3%)	0.535	<0.001	<0.001
Age, median age	44.0 (35.0–50.0)	43.0 (35.0–50.0)	51.0 (45.0–62.0)	51.0 (42.0–58.0)	0.015	<0.001	<0.001
BMI, median BMI	21.6 (19.8–23.7)	21.6 (19.7–23.6)	22.7 (20.7–24.5)	21.9 (20.0–24.1)	0.282	<0.001	<0.001
Abd circumference, median cm	78.5 (73.0–84.3)	78.5 (73.0–84.0)	82.0 (76.0–88.0)	78.9 (73.0–85.0)	0.203	<0.001	<0.001
sBP, median mmHg	112.5 (104.0–121.0)	112.0 (104.0–121.0)	120.0 (111.0–128.0)	115.0 (106.0–123.0)	0.056	<0.001	<0.001
dBP, median mmHg	68.0 (62.0–74.0)	68.0 (62.0–74.0)	74.0 (68.0–80.0)	70.0 (63.0–77.0)	0.038	<0.001	<0.001
Hb levels, median g/dl	14.1 (13.4–14.9)	14.1 (13.4–14.9)	14.4 (13.5–15.2)	14.4 (13.5–15.2)	0.485	<0.001	<0.001
plasma HbA1c levels, median %	5.2 (5.1–5.5)	5.2 (5.1–5.4)	5.4 (5.1–5.7)	5.4 (5.2–5.6)	0.219	<0.001	<0.001
plasma HDL-cholesterol levels, median mg/dl	63.0 (54.0–72.0)	63.0 (54.0–72.0)	61.0 (51.0–71.0)	62.0 (52.0–73.0)	0.595	<0.001	<0.001
plasma LDL-cholesterol levels, median mg/dl	119.8 (103.0–139.0)	119.4 (102.0–139.0)	128.0 (108.0–149.0)	116.0 (97.0–137.0)	0.334	<0.001	<0.001
plasma TG levels, median mg/dl	83.0 (60.0–116.0)	83.0 (60.0–115.0)	96.0 (70.7–140.0)	78.5 (56.0–116.0)	0.261	<0.001	<0.001
plasma AST levels, median IU/l	20.0 (18.0–23.7)	20.0 (17.0–23.5)	21.0 (19.0–26.0)	20.0 (17.0–23.0)	0.355	<0.001	<0.001
plasmaALT levels, median IU/l	18.0 (14.0–25.0)	18.0 (14.0–25.0)	20.0 (15.0–28.0)	17.0 (13.0–25.0)	0.387	<0.001	<0.001
plasma γGTP levels, median IU/dl	24.0 (17.0–36.2)	24.0 (17.0–36.0)	29.0 (19.0–46.0)	22.0 (15.0–35.0)	0.334	<0.001	<0.001
plasma UA levels, median mg/dl	5.4 (4.7–6.2)	5.4 (4.7–6.2)	5.7 (4.9–6.6)	5.3 (4.6–6.0)	0.419	<0.001	<0.001

BMI body mass index, *Abd circum* abdominal circumference, *sBP* systolic blood pressure, *dBP* diastolic blood pressure, *Hb* hemoglobin, *HbA1c* hemoglobin A1c, *HDL* high density lipoprotein, *LDL* low density lipoprotein, *TG* triglyceride, *AST* aspartate aminotransferase, *ALT* alanine aminotransferase, *γGTP* γ-glutamyl transpeptidase, *UA* uric acid

Values are median (interquartile range), and only values of "women" are number (percent)

Statistical testing was performed using pairwise tests, with multiple comparisons adjusted by the Benjamini–Hochberg false discovery rate method

overcome this limitation in LAMP, any hypothesis with a frequency less than σ will not have a p value less than the following level.

$$f(\sigma) = \binom{n_p}{\sigma} \Big/ \binom{N}{\sigma}$$

Here, n_p is the number of objects with positive class labels in D ($n_p < N$). Accordingly, any hypothesis with a frequency less than σ will not be accepted if $f(\sigma) > \alpha/k_D(\sigma)$. Because $f(\sigma)$ is antimonotonic for σ and $\alpha/k_D(\sigma)$ is monotonic, LAMP selects σ^* to balance $f(\sigma^*)$ and $\alpha/k_D(\sigma^*)$. The selected value of σ^* yields the smallest number of candidate hypotheses without applying the tests or missing any significant hypotheses.

For practical considerations, we required each hypothesis to be supported by at least 10 subjects. Because hypotheses involving more than four factors did not satisfy this criterion, the search space was restricted to combinations of up to four factors. This limitation further reduced the number $k_D(\sigma^*)$ of the candidate hypotheses and increased the level $\alpha/k_D(\sigma^*)$ in LAMP. After identifying all significant hypotheses, those whose conditions were supersets of simpler significant hypotheses were excluded, as their additional significance was considered trivial.

Protocol 2: validation of predictive combinations in a large population

In a larger cohort of 2,581,668 general subjects, baseline clinical characteristics were assessed (Table 1), and the occurrence of HT over five years was determined. For each subject, we counted the number of combinations matching the predictive combinations identified in Protocol 1. To test whether HT onset could be predicted using baseline clinical parameters, we evaluated the hypothesis that a greater number of matched predictive combinations at baseline was associated with a higher probability of HT onset during follow-up.

Protocol 3: characterization of combination components

To explore the underlying mechanisms of HT onset, we categorized the components comprising the predictive combinations and quantified the frequency with which each component appeared across all identified combinations.

Statistical analysis

Continuous variables are presented as mean ± standard deviation. In Protocol 1, multiple testing generated by the LAMP procedure was addressed using Bonferroni correction. In Protocol 2, because LAMP evaluates binary indicators of combination matching, Cochran's Q test was applied to assess whether the probability of HT onset increased with the number of matched predictive combinations.

ROC curves were generated using the combination count as a continuous predictor and 5-year incident hypertension as the binary outcome. For each cutoff of the combination count, individuals were classified as predicted positive if their count was greater than or equal to the cutoff; sensitivity and specificity were computed to plot the receiver-operating characteristic (ROC) ROC curve, and the area under the curve (AUC) was calculated using the trapezoidal rule. All statistical tests were two-sided, and a P value < 0.05 was considered statistically significant.

All statistical analyses were performed using Python v310 and packages, including lifelines v0278 (https://github.com/nkimoto/PKMetS).

Results

Table 1 summarizes the baseline clinical characteristics of the cohorts used to evaluate the incidence of HT. Several clinical parameters at baseline differed significantly between individuals who did and did not develop HT during the follow-up period.

Using LAMP, we systematically examined statistically significant combinations among 289 clinical parameters while controlling the familywise error rate by calibrating the Bonferroni correction. This analysis enabled comprehensive characterization of HT outcomes. Among all combinations composed of no more than four clinical parameters, we identified 4802 combinations that were significantly associated with HT onset. Table 2 presents the 23 combinations with the smallest p values, and the complete list of combinations is provided in the Supplemental Table.

To evaluate whether the combinations identified in the discovery cohort of 28,618 individuals were predictive of HT onset, we analyzed an independent cohort of 2,581,668 individuals. We examined whether individuals with a greater number of matching predictive combinations were more likely to develop HT. As the number of predictive combinations applicable to an individual increased, the incidence of HT increased in a stepwise manner over five years ($P < 0.01$; Fig. 1). This result demonstrates a strong and positive association between the number of predictive combinations identified by LAMP and the probability of HT onset during the follow-up period.

Figure 2 presents the ROC curve for HT prediction, demonstrating moderate discriminative performance with AUC of 0.69. These findings indicate that the combinatorial approach provides meaningful predictive information regarding HT onset.

Table 2 The 23 combinations of clinical factors with the smallest p-values

	The combinations of clinical parameters			Adjusted p-value
1	Age: More than 60 years	Uses medication for high blood pressure: No	HDL Cholesterol: More than 40 mg/dL	4.25E-41
2	Age: More than 60 years	HDL Cholesterol: More than 40 mg/dL		5.67E-41
3	Age: More than 60 years	HDL Cholesterol: More than 40 mg/dL	Diastolic Blood Pressure: Less than 90 mmHg	5.67E-41
4	Systolic Blood Pressure: 80–140 mmHg	Age: More than 60 years	HDL Cholesterol: More than 40 mg/dL	1.57E-40
5	Age: More than 60 years	Uses medication for high blood pressure: No		3.52E-39
6	Age: More than 60 years	Uses medication for high blood pressure: No	Diastolic Blood Pressure: Less than 90 mmHg	3.52E-39
7	Age: More than 60 years			4.64E-39
8	Age: More than 60 years	Diastolic Blood Pressure: Less than 90 mmHg		4.64E-39
9	Age: More than 60 years	Plasma AST Levels: Less than or equal to 41 IU/L	HDL Cholesterol: More than 40 mg/dL	5.29E-39
10	Systolic Blood Pressure: 80–140 mmHg	Age: More than 60 years	Uses medication for high blood pressure: No	9.03E-39
11	Plasma ALT Levels: Less than or equal to 41 IU/L	Age: More than 60 years	HDL Cholesterol: More than 40 mg/dL	9.57E-39
12	Systolic Blood Pressure: 80-140 mmHg	Age: More than 60 years		1.19E-38
13	Systolic Blood Pressure: 80–140 mmHg	Age: More than 60 years	Diastolic Blood Pressure: Less than 90 mmHg	1.19E-38
14	Age: More than 60 years	HDL Cholesterol: More than 40 mg/dL	Uses insulin or medication for blood sugar: No	1.05E-37
15	Urinary Glucose: None	Age: More than 60 years	HDL Cholesterol: More than 40 mg/dL	1.14E-37
16	Uses medication for high blood pressure: No	Age: More than 60 years	Plasma AST Levels: Less than or equal to 41 IU/L	2.33E-37
17	Plasma ALT Levels: Less than or equal to 41 IU/L	Age: More than 60 years	Uses medication for high blood pressure: No	2.76E-37
18	Plasma ALT Levels: Less than or equal to 41 IU/L	Systolic Blood Pressure: 80-140 mmHg	Age: More than 60 years	2.91E-37
19	Age: More than 60 years	Plasma AST Levels: Less than or equal to 41 IU/L		3.06E-37
20	Age: More than 60 years	Plasma AST Levels: Less than or equal to 41 IU/L	Diastolic Blood Pressure: Less than 90 mmHg	3.06E-37
21	Plasma ALT Levels: Less than or equal to 41 IU/L	Age: More than 60 years		3.63E-37
22	Plasma ALT Levels: Less than or equal to 41 IU/L	Age: More than 60 years	Diastolic Blood Pressure: Less than 90 mmHg	3.63E-37
23	Age: More than 60 years	Doctor-diagnosed chronic kidney disease or dialysis treatment: No	HDL Cholesterol: More than 40 mg/dL	4.29E-37

BMI body mass index, *Abd* circum abdominal circumference, *sBP* systolic blood pressure, *dBP* diastolic blood pressure, *Hb* hemoglobin, *HbA1c* hemoglobin A1c, *HDL* high density lipoprotein, *LDL* low density lipoprotein, *TG* triglyceride, *AST* aspartate aminotransferase, *ALT* alanine aminotransferase, *γGTP* γ-glutamyl transpeptidase, *UA* uric acid

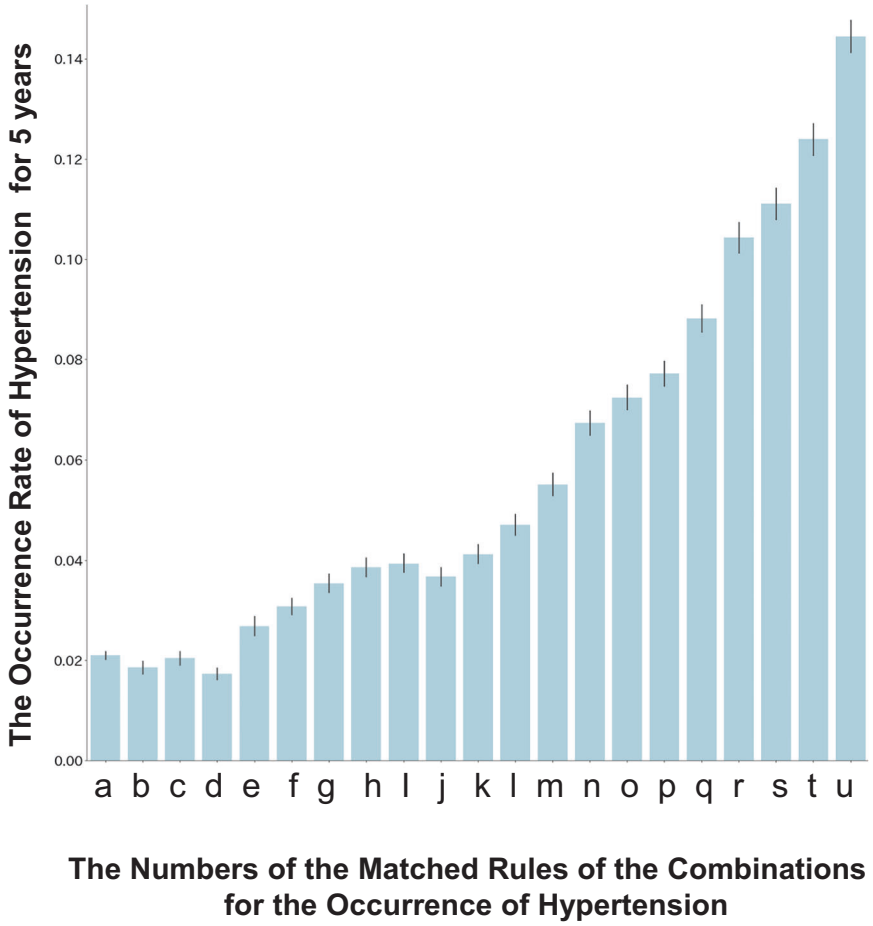


Fig. 1 The relationship between the number of combinations obtained in one cohort and the probability of HT onset in another cohort. As the number of the combinations applied to each subject increases, the probability of the incidence of HT increases. In X axis, a-u indicate a = 0, b = 1-41, c = 42-51, d = 52-60, e = 61-89, f = 90-126, g = 127-183, h = 184-256, i = 257-296, j = 297-324, k = 325-356, l = 357-397, m = 398-459, n = 460-554, o = 555-639, p = 640-711, q = 712-811, r = 812-936, s = 937-1057, t = 1058-1246, and u = 1247-2275

Table 3 summarizes the frequency with which individual clinical factors appeared across the predictive combinations identified by the LAMP analysis (Supplemental Table). This analysis highlights factors that contributed most frequently to combinations associated with HT onset. Commonly represented factors included advanced age, prior medical history, weight gain, prescription of lipid-lowering medications, male sex, elevated low-density lipoprotein cholesterol levels, elevated glucose levels, alcohol consumption, and elevated plasma triglyceride levels, suggesting that these factors may play important roles in the combinatorial pathophysiology of HT.

Discussion

Key messages of the present investigation

The key messages of the present investigation are threefold. First, this study provides evidence supporting the use of an AI-based approach to identify combinations of clinical factors for predicting the onset of HT. For this purpose, we employed a novel data-mining method, LAMP, applied to large-scale clinical data derived from the general population. Second, this study demonstrated that individuals who harbored a greater number of combinations of clinical factors associated with HT onset were more likely to develop HT. This finding reinforces the robustness and clinical relevance of the first observation. Third, we analyzed the frequency with which each factor appeared in the rules generated by the LAMP method, thereby highlighting which clinical parameters play important roles in combinations associated with HT onset. Importantly, the onset of HT appeared to be determined by specific combinations of clinical parameters, even when individual factors were not independently recognized as risk factors for HT.

A novel strategy to identify combinations of medical parameters for HT occurrence

The present study proposes the utility of large-scale data analysis using the LAMP method [15] to identify previously unrecognized rules composed of combinations of clinical factors that predict the probability of HT occurrence. To overcome the problem of combinatorial explosion when searching for meaningful combinations associated with

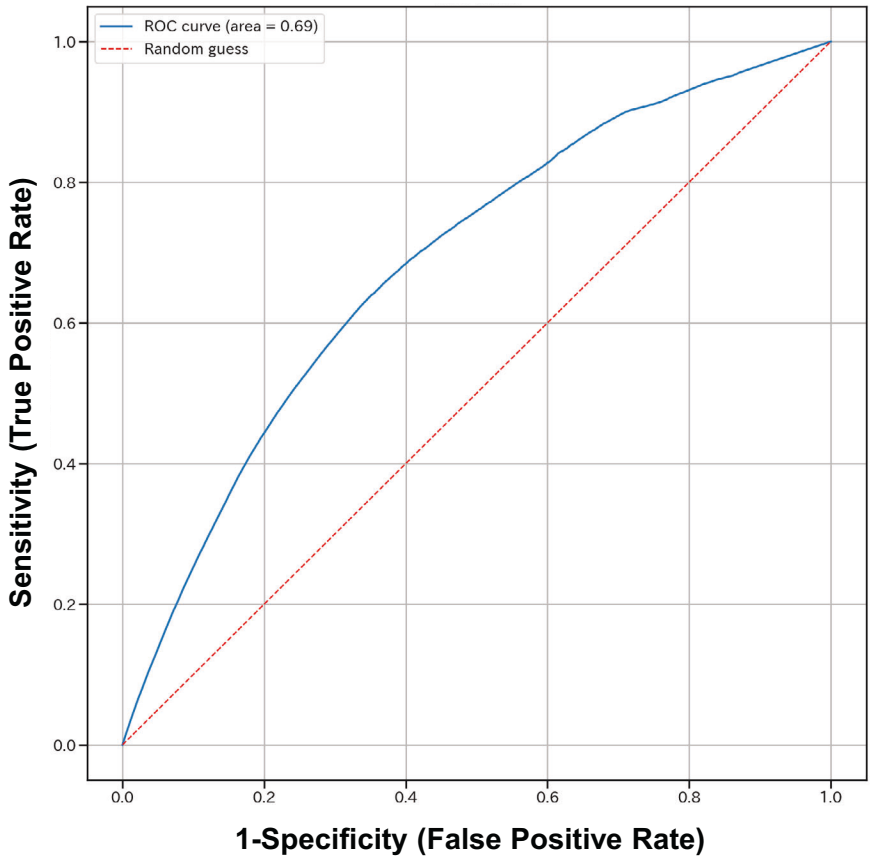


Fig. 2 Receiver operating characteristic (ROC) curve for new-onset HT prediction. Shown is the ROC curve evaluating the discrimination of the predictive summary risk indicator based on the number of hypertension-associated combinations. The area under the curve AUC was 0.69, reflecting moderate discriminative ability. The dashed diagonal line denotes the reference line for random prediction

targeted outcomes, this data-mining approach is well suited to exhaustively uncover combinations of standard clinical parameters that may or may not individually influence HT onset [16, 17]. This strategy enabled us to evaluate not only single parameters but also combinations of parameters obtained from routine health check-ups and medical records that are not obviously or superficially linked to the development of HT.

This data-mining approach shares some conceptual similarities with multivariate analyses; however, conventional multivariate analyses primarily evaluate the independent effects of individual parameters on clinical outcomes and are unable to systematically assess the effects of specific combinations of factors. Moreover, while traditional data-mining approaches often suffer from combinatorial explosion, the LAMP method maintains adequate statistical power even under multiple comparisons and provides statistically significant p values for each rule by appropriately calibrating the Bonferroni correction, thereby minimizing false negatives. Indeed, in the present study, we identified 4,802 combinations with statistically significant p values even after Bonferroni correction.

Importantly, "Age: more than 60 years," is the single determinant identified for HT onset and sging is strongly associated with HT through mechanisms such as atherosclerosis, renal dysfunction, and neural dysregulation (16,17). Because age-related factors were frequently included in the significant combinations, the present findings should be interpreted with the recognition that age is a dominant determinant of new-onset hypertension. Future studies focusing specifically on younger individuals, such as those younger than 60 years, will be needed to identify combinatorial risk profiles that may be more directly useful for early prevention. On the other hand, elevated plasma uric acid levels are also well known to be associated with HT (18,19) and similarly, elevated plasma triglyceride levels contribute to atherosclerosis and are a component of metabolic syndrome, which is closely linked to HT (20). Collectively, these findings in Table 3 suggest that the LAMP-based analytical approach is capable of correctly identifying classical risk factors for HT, thereby supporting the validity of this novel method.

Importantly, because individuals with overt hypertension were excluded at baseline, blood pressure variables in this study reflect variation within the non-hypertensive range rather than established disease.

Mathematical precision models for predicting HT onset

Although we have previously developed an interpretable combinatorial data-mining mathematical model to predict

Table 3 Occurrence frequency of clinical factors in LAMP-derived rules

	Parameters	Count	Parameters	Count	Parameters	Count
1	Age: More than 60 years	1381	Age: 30–60 years	38		
2	Previous Medical History: Yes	593	No Previous Medical History	58		
3	Weight increased by more than 10 kg since age 20: Yes	546	Weight increased by more than 10 kg since age 20:No	50		
4	Uses medication for cholesterol control: Yes	393	Uses medication for cholesterol control: No	184		
5	Gender: Male	363	Gender: Female	31		
6	Plasma LDL Cholesterol: More than 140 and up to 200 mg/dL	359				
7	Fasting Plasma Glucose: More than 126 mg/dL	323	Fasting Plasma Glucose: Less than or equal to 110 mg/dL	125	Fasting Plasma Glucose: More than 110 and up to 126 mg/dL	109
8	Uses medication for high blood pressure: No	320				
9	Diastolic Blood Pressure: Less than 90 mmHg	319				
10	Systolic Blood Pressure: 80-140 mmHg	317				
11	Skips breakfast: No	306				
12	Doctor-diagnosed chronic kidney disease or dialysis treatment: No	296				
13	Doctor-diagnosed anemia: No	271				
14	Doctor-diagnosed stroke or treatment: No	260				
15	HDL Cholesterol: More than 40 mg/dL	256	HDL Cholesterol: Less than or equal to 40 mg/dL	1		
16	Plasma AST Levels: Less than or equal to 41 IU/L	238				
17	Currently a habitual smoker: No	232	Currently a habitual smoker: Yes	1		
18	Uses insulin or medication for blood sugar: No	214	Uses insulin or medication for blood sugar: Yes	43		
19	Eats dinner within 2 hours before sleep: No	213				
20	Urinary Glucose: None	210				
21	Uric Acid: Less than or equal to 7.1 mg/dL	203	Uric Acid: More than 7.1 and up to 9.0 mg/dL	75		
22	Doctor-diagnosed heart disease or treatment: No	203	Doctor-diagnosed heart disease or treatment: Yes	22		
23	Consumes snacks after dinner: No	201				
24	Urinary Protein: None	190	Urinary Protein: Borderline	8		
25	Plasma ALT Levels: Less than or equal to 41 IU/L	187	Plasma ALT Levels: 41-100 IU/L	7		
26	Consumes alcohol: Daily	186	Consumes alcohol: Rarely	43		
27	Plasma Triglycerides: More than 150 and up to 300 mg/dL	174	Plasma Triglycerides: Less than or equal to 150 mg/dL	83		
28	HbA1c Levels: Less than or equal to 6.2%	162	HbA1c Levels: More than 7.0 and up to 8.0%	110		
29	Plasma γ-GTP Levels: Less than or equal to 71 IU/L	158	Plasma γ-GTP Levels: More than 71 IU/L	40		
30	Already improved for more than 6 months	157	Plans to improve lifestyle within 6 months	58	No intention to improve lifestyle	3
31	Walks faster than same-age peers: No	141	Walks faster than same-age peers: Yes	51		
32		136		68		

Table 3 (continued)

	Parameters	Count	Parameters	Count	Parameters	Count
	Performs regular exercise (more than 30 min, more than 2 days/week, more than1 year): No		Performs regular exercise (more than 30 min, more than 2 days/week, more than 1 year): Yes			
33	Experienced weight change of ±3 kg in the past year: No	131	Experienced weight change of ±3 kg in the past year: Yes	32		
34	Performs daily physical activity (≥ 1 hour/day): No	129	Performs daily physical activity (≥ 1 hour/day): Yes	37		
35	Willing to Improve Lifestyle: Yes	94	Willing to Improve Lifestyle: No	58		
36	Gets sufficient rest from sleep: Yes	89	Gets sufficient rest from sleep: No	39		
37	Plasma LDL Cholesterol: Less than or equal to 140 mg/dL	81				
38	Compared to others, eats quickly	65	Compared to others, eats at a normal speed	51	Compared to others, eats slowly	30
39	Alcohol consumption per day: Less than 180 ml	48	Alcohol consumption per day: 180-360 ml	48	Alcohol consumption per day: 360–540 ml	41
40	The combinations of clinical parameters	1				

BMI body mass index, *Abd* circum abdominal circumference, *sBP* systolic blood pressure, *dBP* diastolic blood pressure, *Hb* hemoglobin, *HbA1c* hemoglobin A1c, *HDL* high density lipoprotein, *LDL* low density lipoprotein, *TG* triglyceride, *AST* aspartate aminotransferase, *ALT* alanine aminotransferase, *γGTP* γ-glutamyl transpeptidase, *UA* uric acid

the occurrence of HF (11,21–23), to our knowledge, comprehensive approaches for predicting HT onset that also provide interpretable explanations of the predictive factors remain lacking. Indeed, several previous studies have applied machine learning algorithms to predict the onset of HT using electronic health records or cohort-based risk factors [9, 18–21]. These approaches have primarily aimed to optimize predictive accuracy by modeling individual-level risk through complex, often non-linear functions, such as random forests, gradient boosting, or deep learning architectures. While such models have demonstrated high discrimination performance, they generally operate as black-box predictors and provide limited insight into how specific combinations of clinical parameters jointly contribute to disease onset.

In contrast, the present study adopts a fundamentally different methodological framework. Rather than constructing an individual-level classification model optimized for prediction performance, we employed LAMP to exhaustively identify statistically significant combinations of routine clinical parameters associated with new-onset HT. This combinatorial approach enables transparent identification of interpretable rules, explicitly revealing how multiple factors interact to increase hypertension risk, even when individual components are not independently recognized as risk factors. Importantly, this framework does not rely on model training or parameter tuning, but instead systematically evaluates all possible combinations under strict control of the familywise error rate. This strategy allows comprehensive discovery of clinically meaningful interaction patterns that may be overlooked by conventional machine learning approaches, which tend to prioritize marginal feature importance or overall predictive performance. Furthermore, by quantifying the number of predictive combinations applicable to individuals, we provide a probabilistic and scalable risk stratification scheme that bridges population-level pattern discovery with individual-level risk estimation. Using this framework, we demonstrated that individuals harboring a greater number of HT-associated combinations were more likely to develop hypertension. Importantly, this scenario was validated in an independent large-scale cohort, in which possession of an increasing number of predictive combinations was associated with a stepwise increase in HT incidence over 5 years as demonstrated by the ROC curve shown in Fig. 2. In this context, ROC analysis was not used to evaluate a conventional classification model, but rather to quantify the discriminative capacity of the number of matched hypertension-associated combinations as a summary risk indicator. Accordingly, the observed AUC of 0.69 should be interpreted as reflecting population-level risk stratification performance, rather than individual-level diagnostic accuracy. Notably, this level of discrimination is

comparable to or exceeds that of several established clinical risk scores for incident HT, particularly when applied to population-level risk stratification rather than individual diagnosis.

These findings support the clinical relevance of this combinatorial approach. Nevertheless, whether this framework can prospectively predict future HT onset from a given time point will require validation in forward-looking observational studies.

The significance of the present model for HT onset prediction

HT is known to arise from the interplay of multiple risk factors, including metabolic syndrome, high sodium intake, low potassium intake, obesity, alcohol consumption, physical inactivity, smoking, and unhealthy dietary patterns, which may partly explain regional heterogeneity in HT prevalence [22, 23]. Although such factors were included in several of the identified combinations, our results further indicate that factors not typically regarded as HT risk factors—such as stable body weight since early adulthood, normal fasting glucose levels, absence of diabetes mellitus, normal uric acid levels, and preserved liver function—may nonetheless contribute to HT-associated combinations when interacting with other clinical parameters. These findings should not be interpreted as protective effects of "normal" values per se, but rather as context-dependent components within higher-order interaction patterns.

While the biological significance of such combinations may not be readily explained by current reductionist or functional analyses, these findings underscore the multifactorial complexity of HT pathogenesis and suggest that future studies may uncover previously unrecognized mechanisms underlying HT development.

Limitations of the present study

Several limitations of the present study should be acknowledged. First, the definition of HT onset was based on ICD-10 diagnostic codes, and concerns regarding the accuracy of ICD-based diagnoses may be raised. However, in Japan, HT has long been highly prevalent and closely monitored due to the historically high incidence of cerebrovascular disease. Consequently, guidance from the Japanese Society of Hypertension is well established, and both specialists and general practitioners are likely to diagnose HT in a rigorous and standardized manner, reducing the likelihood of substantial diagnostic misclassification.

Second, the cohort analyzed in this study consisted exclusively of Japanese individuals, raising questions regarding the generalizability of the findings to other Asian populations or to Western populations. Nevertheless, numerous prior studies have demonstrated that the etiology, pathophysiology, and pharmacological responses of HT do not differ substantially between Japanese and non-Japanese populations. Therefore, it is plausible that the findings obtained in this Japanese cohort may be extrapolated to broader populations worldwide, although validation studies in other countries will be required.

Third, information on dietary salt intake was not available in the present dataset. Because high salt intake is an established risk factor for HT, its potential contribution to new-onset HT and to the identified predictive combinations could not be evaluated in this study. Future studies incorporating dietary information, including salt intake, will be needed to further clarify the relationship between lifestyle factors and combinatorial risk profiles for HT onset.

Fourth, replication of this study would require the acquisition of a large number of clinical parameters and long-term follow-up over several years. As a next step, we plan to conduct such prospective studies to further validate and refine this analytical framework.

Conclusion

In conclusion, we identified specific combinations of clinical parameters that predict new-onset HT in the general population, and demonstrated that a greater number of such combinations proportionally increased the probability of HT onset. This quantitative AI-based approach may be useful for stratifying HT risk and identifying high-risk individuals for new-onset HT in the general population.

Acknowledgements There is no person to be acknowledged in preparation of this manuscript.

Funding This study was funded by the Japan Heart Foundation in Japan.

Compliance with ethical standards

Conflict of interest The conflict of interests of all the authors are as follows. Relationships to industry do not exist for Y.M., N.K, K.O., Y.Y., T.A., M.Y., T.W., and S.T. M.K. reports personal fees from Daiichi-sankyo, personal fees from Viatris, grants and personal fees from Ono, grants from Novartis, grants and personal fees from Tanabe-mitubishi, grants from Takeda, grants and personal fees from Astra Zeneca, grants and personal fees from Boehringer-ingelheim, grants from Kowa, personal fees from Otsuka, personal fees from Eli Lilly outside the submitted work.

References

1. Levy D, Larson MG, Vasan RS, Kannel WB, Ho KK. The progression from hypertension to congestive heart failure. JAMA. 1996;275:1557–62.

2. Ho KK, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Study. J Am Coll Cardiol. 1993;22:6a–13a.
3. Drazner MH. The progression of hypertensive heart disease. Circulation. 2011;123:327–34.
4. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA*. 2002;288:2981-97.
5. Ettehad D, Emdin CA, Kiran A, Anderson SG, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. Lancet. 2016;387:957–67.
6. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol. 2018;71:e127–e248.
7. Whelton PK, Carey RM. The 2017 clinical practice guideline for high blood pressure. JAMA. 2017;318:2073–4.
8. Carretero OA, Oparil S. Essential hypertension: part I: definition and etiology. Circulation. 2000;101:329–35.
9. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine learning for hypertension prediction: a systematic review. Curr Hypertens Rep. 2022;24:523–33.
10. Terada A, Okada-Hatakeyama M, Tsuda K, Sese J. Statistical significance of combinatorial regulations. Proc Natl Acad Sci. 2013;110:12996–3001.
11. Miyashita Y, Hitsumoto T, Fukuda H, Kim J, Washio T, Kitakaze M. Predicting heart failure onset in the general population using a novel data-mining artificial intelligence method. Sci Rep. 2023;13:4352.
12. Fukuda H, Shindo K, Sakamoto M, Ide T, Kinugawa S, Fukushima A, et al. Elucidation of the strongest predictors of cardiovascular events in patients with heart failure. EBioMedicine. 2018;33:185–95.
13. Umemura S, Arima H, Arima S, Asayama K, Dohi Y, Hirooka Y, et al. The Japanese Society of Hypertension Guidelines for the Management of Hypertension (JSH 2019). Hypertens Res. 2019;42:1235–481.
14. Ohya Y, Sakima A, Arima H, Fukami A, Furuhashi M, Ishida M, et al. Key highlights of the Japanese Society of Hypertension Guidelines for the management of elevated blood pressure and hypertension 2025 (JSH2025). Hypertens Res. 2025;48:2500–11.
15. Terada A, Okada-Hatakeyama M, Tsuda K, Sese J. Statistical significance of combinatorial regulations. Proc Natl Acad Sci USA. 2013;110:12996–3001.
16. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst. 2002;26:445–63.
17. Kim J, Washio T, Yamagishi M, Yasumura Y, Nakatani S, Hashimura K, et al. A novel data mining approach to the identification of effective drugs or combinations for targeted endpoints-application to chronic heart failure as a new form of evidence-based medicine. Cardiovasc Drugs Ther. 2004;18:483–9.
18. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J Med Int Res. 2018;20:e22.
19. Hwang SH, Lee H, Lee JH, Lee M, Koyanagi A, Smith L, et al. Machine Learning-Based Prediction for Incident Hypertension Based on Regular Health Checkup Data: Derivation and Validation in 2 Independent Nationwide Cohorts in South Korea and Japan. J Med Int Res. 2024;26:e52794.
20. Schjerven FE, Ingeström EML, Steinsland I, Lindseth F. Development of risk models of incident hypertension using machine learning on the HUNT study data. Sci Rep. 2024;14:5609.
21. Islam SMS, Talukder A, Awal MA, Siddiqui MMU, Ahamad MM, Ahammed B, et al. Machine learning approaches for predicting hypertension and its associated factors using population-level data from three South Asian Countries. Front Cardiovasc Med. 2022;9:839379.
22. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nat Rev Nephrol. 2020;16:223–37.
23. Beilin L, Puddey I, Burke V. Lifestyle and hypertension. Am J Hypertens. 1999;12:934–45.