Creating a General-Purpose Generative Model for Healthcare Data based on Multiple Clinical Studies

Hiroshi Maruyama[1,2,3†✉], **Kotatsu Bito**[4†], **Yuki Saito**[4], **Masanobu Hibi**[5], **Shun Katada**[5], **Aya Kawakami**[4], **Kenta Oono**[2], **Nontawat Charoenphakdee**[2], **Zhengyan Gao**[2], **Hideyoshi Igata**[2], **Masashi Yoshikawa**[2], **Yoshiaki Ota**[2], **Hiroki Okui**[2], **Kei Akita**[2], **Shoichiro Yamaguchi**[2], **Yohei Sugawara**[2], **Shin-ichi Maeda**[2]

[1]Kao Corporation; [2]Preferred Networks Inc.; [3]Research into Artifacts, Center for Engineering, The University of Tokyo; [4]Digital Business Creation, Kao Corporation; [5]Human Healthcare Research Laboratories, Kao Corporation

✉ **For correspondence:** maruyama@acm.org

[†]These authors contributed equally to this work.

Present address: Digital Business Creation, Kao Corporation, 1-14-10 Nihonbashi Kayabacho, Chuo-ku, Tokyo, Japan

Data availability: Owing to ethical and contractual restrictions, the underlying individual-level data cannot be made publicly available. Researchers who wish to request access to the data should contact the Digital Business Creation division at Kao Corporation (kaodbc-contact@kao.com). Requests will be considered in consultation with the relevant ethical committees and IRBs, and access may be granted subject to appropriate agreements and approvals. The statistical properties of the data are approximated in the model and are available via an application program interface (API).

Funding: This study was solely sponsored by Kao Corporation (Tokyo, Japan), which provided full financial support for the entire research project. Kao Corporation covered all costs associated with project management, data collection, and computational resources for model development of Preferred Networks, Inc. (Tokyo, Japan), which was commissioned as a model development partner. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. (Funder website: https://www.kao.com/) None of the authors have received any other specific funding for this study.

Competing interests: KB, YSaito, MS, SK, and AK are employees of Kao Corporation (Tokyo, Japan). KO, NC, ZG, HI, MY, YO, HO, KA, SY, Y Sugawara, and SM are employees of Preferred Networks, Inc. (Tokyo, Japan). HM was an executive fellow of Kao Corporation and a senior advisor of Preferred Networks, Inc. The authors have no other competing interests to declare.

Abstract

Data for healthcare applications are typically customized for specific purposes but are often difficult to access due to high costs and privacy concerns. Rather than prepare separate datasets for individual applications, we propose a novel approach: building a general-purpose generative model applicable to virtually any type of healthcare application. This generative model encompasses a broad range of human attributes, including age, sex, anthropometric measurements, blood components, physical performance metrics, and numerous healthcare-related questionnaire responses. To achieve this goal, we integrated the results of multiple clinical studies into a unified training dataset and developed a generative model to replicate its characteristics. The model can estimate missing attribute values from known attribute values and generate synthetic datasets for various applications. Our analysis confirmed that the model captures key statistical properties of the training dataset, including univariate distributions and bivariate relationships. We demonstrate the model's practical utility through multiple real-world applications, illustrating its potential impact on predictive, preventive, and personalized medicine.

Introduction

Advances in information technology, particularly in machine-learning and sensing technologies, are revolutionizing human healthcare by enabling continuous monitoring and analysis of individual health status [Alhussein and Muhammad, 2018; Artico, Edge III, and Langham, 2022; Čolaković and Hadžialić, 2018; Lu et al., 2021; Pradhan, Bhattacharyya, and Pal, 2021]. Digital data collected through various measurements can be integrated to create comprehensive health profiles, allowing for early detection of health risks and personalized interventions. This data-driven approach is crucial for realizing predictive, preventive, and personalized medicine [Hsiao et al., 2022; Wang et al., 2021]. For example, by analyzing patterns in lifestyle, diet, and physical activity data, healthcare providers can develop tailored interventions that address individual risk factors before they

lead to serious health conditions.

A major challenge in developing such healthcare solutions, however, is the limited availability of comprehensive health data. Data collection is often hindered by high costs, privacy concerns, and the fragmented nature of health records. Traditional approaches require collecting data specific for each application, making it inefficient and sometimes impractical to develop multiple healthcare solutions.

To address these challenges, we propose the Virtual Human Generative Model (VHGM), a novel statistical framework that can estimate missing attribute values from known attribute values and generate synthetic but realistic human health data. Unlike conventional statistical models, the VHGM can:

1. Capture complex relationships among over 2000 diverse health attributes
2. Estimate missing attribute values from known attribute values
3. Generate synthetic data that preserves the statistical properties of real populations
4. Support multiple healthcare applications through a single model

There are two key innovations in our approach. One is the integration of multiple independent data sources to create a high-dimensional training dataset. While existing clinical studies typically include fewer than 100 attributes [Pezoulas et al., 2024], limiting their applicability, our method combines data from many sources using statistical linking techniques, without relying on personally identifiable information.

The other is maintaining the quality of the VHGM. Current deep learning-based machine learning is stochastic and there is no guarantee of the "correctness" of the model outputs. Among several ongoing technical discussions related to improving the trustworthiness of a model, some focus on augmenting the training datasets [Kreuzberger, Kühl, and Hirschl, 2023] and others focus on the characteristics of the deep neural network [Maruyama, 2018]. We take a practical approach to the VHGM, combining a robust set of quality metrics to objectively measure the model quality with a transparent governing process with multiple stakeholders.

The primary contributions of this paper are:

1. Development of a novel method for combining data from multiple clinical studies while preserving their statistical relationships
2. Implementation of the VHGM, a generative model with more than 2000 heterogeneous health attributes across diverse categories, together with a practical quality assurance process
3. Demonstration of the VHGM's practical utility through multiple real-world healthcare applications

Materials & methods

Design of the VHGM

The VHGM is produced and operated by three data-processing steps and one governing process (see Fig 1). Step 1 is to prepare the data. We used three data sources, each of which was comprised of one or more table-structured datasets. Data Source A was specifically prepared for the VHGM to obtain diverse health attributes simultaneously from approximately 1000 participants. Data Source B was commercially available data on annual health checkups and health insurance receipts of over one million individuals. Re-purposing of these data conforms to the Japanese Privacy Law and was approved by the Information Security Committee of Kao Corporation. Data Source C was a collection of previously reported studies to supplement specific health attributes. These studies were conducted internally by Kao Corporation (Tokyo, Japan), each of which was individually approved for that particular study. Re-purposing these datasets for the VHGM is covered by an umbrella approval in April 2021 by the IRB of the Kao Corporation and the Preferred Networks, Inc (Tokyo, Japan).

Step 2 is to integrate all the data sources into a single training dataset. The attributes to be extracted from each data source were determined by the *model schema*. The model schema also determines the data type of each attribute.

Step 3 is to train the generative model [Oono et al., 2023]. The resulting VHGM model was deployed as a commercial Application Program Interface (API) service.

During the whole process, the governing committee oversaw the quality control of the VHGM.

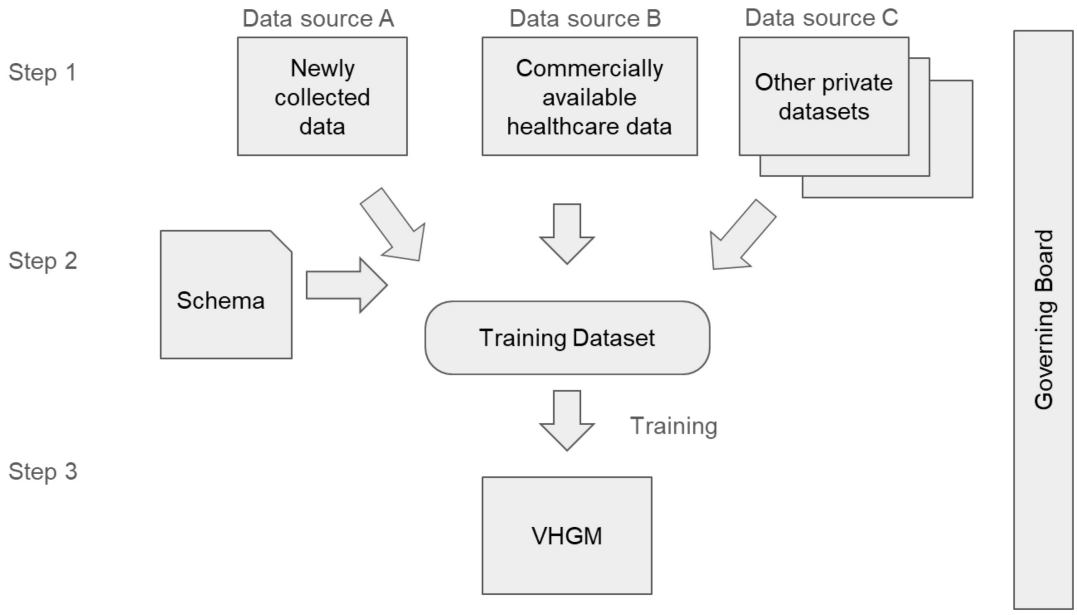

Figure 1. Schematic diagram of the VHGM development and operation. Data obtained from Data Sources A, B, and C are combined into a single training dataset to build a general-purpose statistical model called the VHGM, which is accessible via the API.

Data Source A

Data Source A is from a single-center cross-sectional observational study that was conducted with adult men and women living in narrowly defined metropolitan areas of Japan (i.e., Tokyo, Kanagawa, Chiba, Saitama, Ibaragi, Tochigi, and Gunma prefectures). All measurements were performed by trained research coordinators and medical doctors using standard operating procedures during two outpatient visits to the Ueno Asagao Clinic (Tokyo, Japan), 1 week apart. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines were applied according to the study objectives [Von Elm et al., 2014]. The study protocol is available online [Hibi et al., 2023].

Ethics approval, informed consent, and participation

The study was approved in October 2021 by the IRB of Kao Corporation (Tokyo, Japan; approval #K0023-2108) and Preferred Networks, Inc (Tokyo, Japan; approval #ET22110047). Eligibility was evaluated by asking potential participants a few questions. All participants provided written informed consent to participate in the study. The consent form explains in detail which data would be used in the study and obtains consent for the use of anonymized data. It also stated that statistical models developed through the use of participants' anonymized data may be used in the future by Kao Corporation or its commissioned contractors. The study in Data Source A was registered at the University Hospital Medical Information Network (UMIN; UMIN000045746) on October 14, 2021. Recruitment was started on October 19, 2021, and ended on February 25, 2022.

Participants and eligibility

Eligible participants were consecutively recruited over a 5-month period from October 2021 through February 2022. The participants were recruited via a website administered by TES Holdings (Tokyo, Japan). Participants were stratified into age groups by decade (20-29, 30-39, 40-49, 50-59, 60-69, and $\geq$ 70 years) to match the decade ratio of the typical adult Japanese population. The major inclusion criteria were as follows: (1) Japanese men and women aged $\geq$ 20 years and (2) individuals able to complete the questionnaires and surveys. Major exclusion criteria were as follows: (1) individuals undergoing hospitalization for serious diseases (e.g., diabetes, hypertension, arteriosclerosis, heart disease, malignancy, Alzheimer's disease, etc.), (2) individuals who could not come to the outpatient unit by themselves, and (3) individuals with dementia or suspected dementia. The detailed inclusion and exclusion criteria are available in the protocol [Hibi et al., 2023].

Measurements and data processing

Numerous health attributes across diverse categories were collected. The parameters were grouped into the following 16 measurement categories: blood pressure and arterial stiffness, lifestyle investigation and questionnaire, cognitive function analysis, laboratory analysis, oral glucose tolerance test, anthropometric measurements, skin surface spectroscopy, physical performance tests, hand surface analysis, liquid chromatography-tandem mass spectrometry, body odor analysis, lipids in the stratum corneum and sebum analysis, hair loss determination, lipid mediator analysis, skin surface lipids (SSL)-RNA analysis, and microbiota analysis, based on the measurement methods described in the protocol [Hibi et al., 2023]. Details on the SSL-RNA, intestinal microbiota, and saliva microbiota are described in S1 Text. The data management details are also described in the protocol paper [Hibi et al., 2023].

Data analysis

We reviewed basic statistical characteristics of the data and compared them with recent official statistics in Japan. A correlation matrix using the Spearman rank correlation was generated to overview relationships between attributes and data sparseness. To create the correlation matrix, only real, positive, ordered categorical, and binary (only two selective in categorical type) attributes were used.

Data Source B

In Japan, all employers are required to provide healthcare insurance coverage for their employees through company-specific insurance associations. Under legislation enacted by the Japanese government, healthcare-related records can be utilized for research and development purposes without individual consent, provided they undergo an approved anonymization process [Ministry of Justice, 2017]. Several private-sector data aggregators make such anonymized data commercially available. Data Source B is comprised of two comprehensive datasets covering the same set of approximately one million individuals, including both employees and their dependents in Japan in 2019:

1. Annual health examination records, including:
    - Physical measurements (height, weight, *etc.*)
    - Blood test results
    - Responses to health-related lifestyle questionnaires
2. Medical and dental consultation records, including:
    - Diagnosed conditions
    - Diagnostic tests and treatments performed
    - Drug prescriptions administered
    - Insurance points (used for calculating reimbursement amounts)

We preprocessed the purchased data, selecting 56 attributes from the annual health examination records and extracting 199 attributes for major disease diagnosed, major test and treatment procedures performed, or major drugs administered, each of which represents how many times the individual visited the doctor for that particular disease, procedure, or prescription drug during the year. In addition, we also added three extra attributes: one for the total insurance points of the year (roughly representing how much money the individual spent on medical services in the year), one for the insurance points related to medical (non-dental) services, and one for the insurance points related to dental services. Furthermore, we introduced three binary flag attributes representing service utilization: one indicating whether the individual ever visited a medical (non-dental) doctor in the year, one indicating whether the individual ever visited a dentist in the year, and one indicating whether the individual used both medical and dental services in the year.

Data Source C

The collection criteria of previous studies in Kao Corporation were as follows: 1. an adequate number of participants for modeling (≥ 100 participants in each clinical study), 2. inclusion of common attributes (age, sex, height, weight, *etc.*), and 3. gender balance (i.e., exclusion of datasets containing only male or only female participants). Under these criteria, we selected one cross-sectional study on visceral fat accumulation [Takase, Sakane, et al., 2019] as Data Source C-1 and 12 intervention clinical trials on drinks including green tea catechins and coffee chlorogenic acids [Chikama et al., 2006; Kozuma et al., 2005; Matsui, Kinoshita, et al., 2018; Matsui, Takeshita, et al., 2016; Nagao, Hase, and Tokimitsu, 2007; Nagao, Ochiai, Katsuragi, et al., 2007; Nagao, Ochiai, Watanabe, et al., 2009; Takase, Nagao, et al., 2008; Takeshita et al., 2008; Tsuchida, Itakura, and Nakamura, 2002; Watanabe et al., 2019; Yamaguchi et al., 2007] as Data Source C-2. The cross-sectional study of Data Source C-1 includes basic measurements such as general blood testing and lifestyle questionnaires. The intervention trials of Data Source C-2 include basic measurements (e.g., weight, height, and blood pressure) and specialized measurements (e.g., visceral fat area, lipid profile, and gastrointestinal hormones) to assess metabolic syndromes. Data from the first visit before interventions and the subsequent changes with active ingredients after interventions were extracted and treated as baseline participant characteristics at a single reference time point.

Design of model schema

The VHGM represents a joint probability distribution over a set of random variables $X_1, X_2, ..., X_k$, where each random variable represents an *attribute* of human health data. The *model schema* defines which attributes to extract from the data sources, along with their domains and semantic interpretations. While maximizing the number of attributes could potentially increase the model's utility for future applications, including superfluous attributes that are rarely used or contribute minimally to the estimation of other attributes can adversely affect both model accuracy and computational efficiency. Therefore, we established the following criteria for attribute selection:

1. Missing rate: Attributes with high missing data rates are excluded as they lead to less reliable estimations
2. Potential utility: Attributes with limited applicability in anticipated future applications are omitted
3. Independence: Attributes showing minimal correlation with other variables are less valuable for joint estimation and may be excluded

The model is designed to handle *heterogeneous* data types, accommodating various attribute distributions and domains. We assume that each type has a parametric distribution, such as the Gaussian distribution. Based on our training algorithm requirements [Oono et al., 2023], we categorized attributes into the following types:

1. *Real*: Continuous variables following a normal distribution (e.g., height)

2. *Positive*: Strictly positive continuous variables following a log-normal distribution (e.g., blood glucose levels)
3. *Count*: Discrete variables following a Poisson distribution (e.g., number of doctor visits)
4. *Categorical*: Nominal variables with a finite set of unordered options (e.g., sex)
5. *Ordered Categorical*: Categorical variables with inherent ordering (e.g., drinking habit as in *Never, Sometimes, Everyday*)

Model algorithm

Combining multiple datasets

Privacy Preserving Record Linkage Systems are tools designed to link records across multiple datasets while protecting individual privacy [Pathak et al., 2024]. These systems typically require access to personally identifiable information before the de-identification process and assume the existence of a sufficient number of common subjects across datasets. The VHGM training algorithm [Oono et al., 2023] takes a fundamentally different approach, eliminating the need for personally identifiable information or common subjects across datasets. Instead, it employs statistical linkage, leveraging common attributes (such as age and sex) that naturally occur across different datasets without requiring shared identifiers.

The statistical linkage in the VHGM can be conceptually expressed through transitional conditional probability (implementations details are given in [Oono et al., 2023]). To establish an "indirect relation" between variable X in dataset 1 and variable Y in dataset 2, one may:

1. Estimate the conditional probability distribution $P(Z|X)$ in dataset 1
2. Estimate the conditional probability distribution $P(Y|Z)$ in dataset 2
3. Apply the marginalization rule of conditional probability to calculate

$$P(Y|X) = \int P(Y|Z,X)P(Z|X)dz$$

   where Z represents common variables present in both datasets. Here, we assume $P(Y|Z) = P(Y|Z,X)$, i.e., given Z, Y is independent of X.

The process of combining these heterogeneous datasets into a unified training dataset is illustrated in Fig. 2. We employed a row-wise concatenation approach, where:

- Each record from the source datasets becomes a separate record in the combined dataset
- Attributes not present in a particular source dataset are treated as missing values
- Common attributes across datasets (e.g., age and sex) serve as implicit linking features

Model architecture

Given the systematic nature of missing values in our combined dataset (as opposed to random missingness), we designed a model architecture that is robust in the presence of large-scale structured missing data patterns [Oono et al., 2023]. The main idea is inspired by Vision Transformer (ViT) [Dosovitskiy et al., 2020] for image recognition where an input image is split into a set of image "patches," and a Transformer is used to capture the semantic relationships among these patches.

Similarly, instead of treating the input as a fixed-size vector in many table-based machine learning systems, our algorithm takes a sequence of "tokens" of observed (non-missing) attributes as if it were a sequence of words. These tokens are embedded into a fixed-dimensional space, accommodating various encoding schemes such as one-hot encoding for categorical attributes.

Figure 3 illustrates the architecture of our model. Our transformer-based encoder leverages attention mechanisms to capture the relationships among the observed input tokens and transforms it into a sequence of latent representations (blue boxes in the diagram). Missing attributes do not contribute during this encoding process. Subsequently, the latent representation for all attribute is constructed by combining the transformed tokens for observed attributes and a default learnable token assigned to each missing attribute (yellow boxes in the diagram). From this

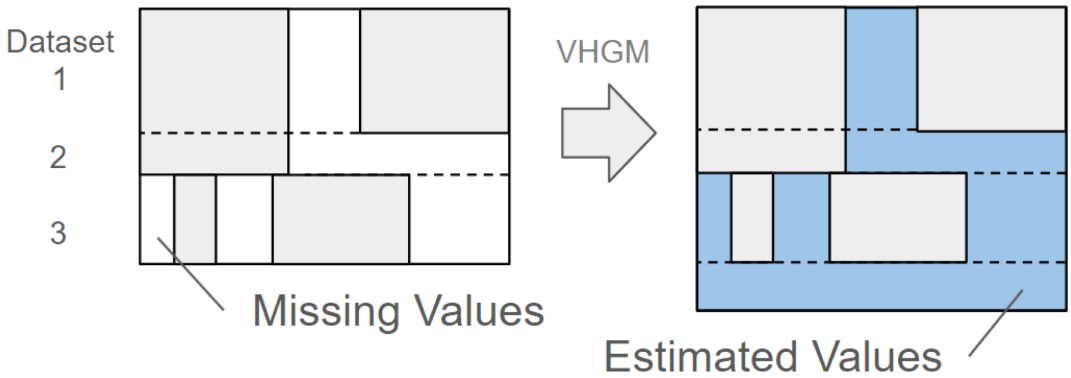

Figure 2. Dataset concatenation and data generation. The diagram shows the relationship between records (rows) and attributes (columns), with distinct datasets represented as row blocks. White areas indicate missing values and blue areas represent estimated values. Additional rows on the right-hand side indicate new datasets generated by the model.

unified latent representation, the decoder generates the estimated distribution of every attribute. The output of each attribute is a set of estimated parameters for the attribute, depending on its type (e.g., mean and standard deviation for *real* attribute modeled as Gaussian).

Model training

We employed two techniques to train the model. The first is Masked Modeling [He et al., 2022] in which certain input attributes are intentionally masked, and the model is trained to recreate their original values. This approach is particularly effective for the missing-value imputation task targeted by the VHGM.

The other is a *two-stage training approach*: In the first stage, the model is trained separately on individual datasets while in the second stage the model is fine-tuned by the combined dataset. We found that this two-stage training strategy significantly reduced training time without noticeable degradation in accuracy.

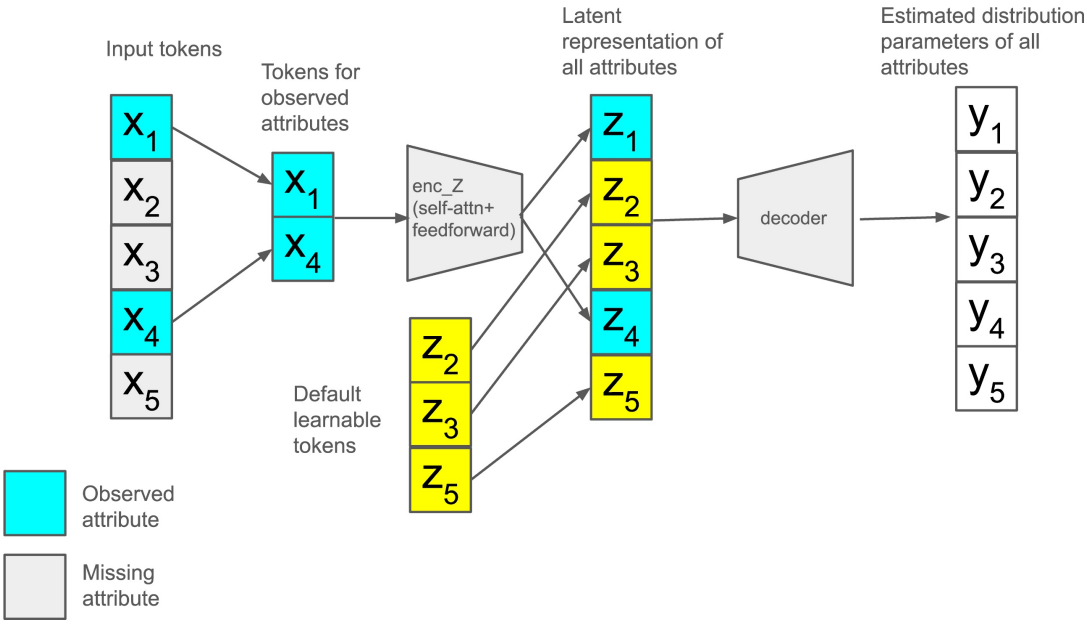

Figure 3. Model architecture. Tokens for observed attributes (blue) are transformed into latent representations and combined with a default learnable token for missing attributes (yellow).

Model quality

The VHGM aims to approximate the joint probability distribution across more than 2000 attributes. Therefore, its quality should be evaluated based on how accurately it captures the underlying real-world distribution. This evaluation presents three significant technical challenges:

1. **The challenge of ground truth:** The actual real-world distribution is unknown and cannot be precisely estimated from observed data. To address this, we employed the standard approach of data splitting:
   - For each of the data sources, randomly select 10% of the original records as a holdout set, maintaining the sex and age distributions
   - Used the remaining 90% for model training
   - Maintained strict isolation of the holdout set from model developers to prevent information leakage and ensure unbiased evaluation
   - Evaluated imputation errors using the holdout set

   Using the 10% holdout sets, we calculated the imputation errors for each attribute in each data source. The value of a target attribute in a record from the holdout set was masked, and an imputed value was obtained using 10 attributes in the same record as input for the VGHM. These 10 attributes were chosen based on their strong relationship with the target attribute using the VGHM. Errors were calculated by comparing the target attribute value with the imputed value. The method details, depending on the data type of the target attribute, are described in S2 Text.
2. **The challenge of high dimensionality:** Evaluating the similarity between high-dimensional joint distributions is inherently complex. In the absence of standardized methods for such evaluation, we developed a practical three-component assessment framework:
   (a) *Univariate analysis:* For each attribute, we compared the marginal distributions between the training dataset and model output. For real type attributes, we evaluated the overlapping area between the histograms of the training dataset and model output. Detailed methods are described in S3 Text.
   (b) *Bivariate analysis:* For 70 pre-selected attribute pairs $\langle X, Y \rangle$ that exhibited high correlations in the training dataset, we compared the conditional distributions $P(Y|X)$ between the training data and model output. Detailed methods are described in S3 Text.
   (c) *Scenario-based analysis:* In this study, “scenario-based analysis" refers to an evaluation approach in which the VHGM is tested under predefined conditions that represent realistic or hypothetical use cases relevant to potential applications. In this analysis, outputs were observed in the model response to pre-selected inputs based on anticipated use case scenarios. Given that there may not always be sufficient records in the training data to validate the same combinations of input values, we assessed the direction and magnitude of changes in the obtained results to ensure they were intuitively consistent and comparable to prior knowledge.
3. **The challenge of validation with external datasets:** Validating a generative model against external datasets poses substantial challenges due to variations in data collection protocols and population demographics. To assess the model’s generalizability, we conducted a comparative analysis using two well-established, independent datasets: the National Health and Nutrition Survey, available via the Portal Site of Official Statistics of Japan (e-Stat) [The National Statistics Center, 2021b], and the U.S. National Health and Nutrition Examination Survey (NHANES) [Centers for Disease Control and Prevention, National Center for Health Statistics, 2021].
   The validation focused on 24 nutritional intake attributes (e.g., calories, protein) that were derived from diet record or recall methods and determined to exhibit high semantic equivalence across all three datasets (see S4 Text). For each attribute, we statistically compared the dis-

tributions to evaluate concordance. Specifically, we treated the mean value from the VHGM output as a sample mean and calculated its z-score relative to the corresponding distribution in each external dataset (e-Stat or NHANES). Assuming normality, we further computed the overlapping area between the VHGM distribution and each external dataset, providing a quantitative measure of distributional similarity.

In addition to these evaluation components, we also conducted a benchmark comparison with widely used tabular generative models, including TVAE, CTGAN [Xu et al., 2019], and Gaussian Copula. As a common performance metric for this comparison, we adopted *Machine Learning Efficiency*, which evaluates whether synthetic data can give rise to a machine learning model with performance comparable to that trained on the original data. Because not all baseline models can handle discrete variables, we employed a regression task.

This evaluation approach aligns with similar metrics used in recent synthetic healthcare data generation efforts [Pezoulas et al., 2024]. Previous studies similarly emphasized the importance of measuring the fidelity of synthetic data across multiple dimensions of analysis.

Governance process

In this study, the term "governance process" refers to the structured set of policies, decision-making mechanisms, and oversight activities that define stakeholder roles, guide schema updates and new model releases, and manage risks to ensure the secure, ethical, and effective operation of the VHGM. Our governance framework addresses the needs and concerns of four key stakeholder groups:

1. **Data subjects**: Individuals whose health data contribute to the model. Their primary concerns are data privacy and protection, which are mitigated through robust anonymization protocols.
2. **Data owners**: Organizations providing source datasets. Their primary concerns are data security and protection of intellectual property, which are mitigated through contractual agreements, system security, and usage monitoring.
3. **Application owners**: Organizations providing healthcare solutions using the VHGM. Their primary concerns are model reliability, availability, and performance, which are managed by support through technical documentation and service-level agreements.
4. **End users**: Consumers of VHGM-based applications. Their primary concerns are trustworthiness and validity of the VHGM outputs, which are addressed through transparent validation processes and a clear limitations disclosure.

To ensure effective oversight, we established a multi-stakeholder governance committee that (1) conducts monthly meetings to review operations, (2) makes final decisions on the model schema, (3) approves new model releases, and (4) evaluates potential risks and determines mitigation strategies. Additionally, the committee members are carefully nominated to cover diverse backgrounds, including life science researchers, clinical study experts, data scientists, and marketers with expertise in healthcare applications, thereby ensuring that users' perspectives, healthcare demands, and ethical and privacy matters are appropriately considered. The terms and conditions of the VHGM API service are also carefully designed to ensure governance in providing healthcare solutions using the VHGM.

During this whole process, we maintained transparency through open communication with multiple channels as follows:

- Publication of technical papers on study planning [Hibi et al., 2023], outcomes (this paper), and training algorithms [Oono et al., 2023]
- Monthly newsletters [Kao Corporation, 2023]
- Clear documentation of model capabilities and limitations

Machine learning models can never be perfect and require continuous refinement. As such, we periodically release newer versions of the VHGM model. This process is often referred to as "MLOps" [Kreuzberger, Kühl, and Hirschl, 2023], and is known to be complex because improvements in certain aspects may affect others. Through the transparent governance process described above, the VHGM allows stakeholders to make informed decisions about which model version best suits their specific needs.

Results

This section describes the results for the latest model of the VHGM as of August 2025, named *pollux* (see Table 11).

Data Source A

A total of 997 participants were included in the study and their data were obtained. Three of the participants placed restrictions on the use of their data and did not consent to secondary use of their data. Thus, the data from 994 participants were used for data characterization, model development, and model applications. Figure 4 summarizes the participants' flow process.

Participants

The sex and age ratios of participants in these visits closely mirror those of the adult Japanese population (see the details in S1 Table and S2 Table). Table 1 shows the characteristics of the participants. These values are consistent with recent official statistics provided by the Japanese government [The National Statistics Center, 2021a]. Thus, this dataset may approximately represent the statistical characteristics of the Japanese population.

Table 1. Data Source A participant characteristics.

Characteristics	Male	Female	Male (Ref.)	Female (Ref.)
Age (y), mean (SD)	51.36 (16.27)	51.16 (16.51)	n/a	n/a
Height (cm), mean (SD)	169.80 (5.83)	157.19 (5.74)	167.7 (6.9)	154.3 (6.7)
Weight (kg), mean (SD)	68.88 (12.42)	52.79 (9.76)	67.4 (12.0)	53.6 (9.2)
BMI (kg/m^2), mean (SD)	23.87 (3.96)	21.35 (3.69)	23.9 (3.6)	22.5 (3.7)

Mean (standard deviation [SD]) of height, weight, and BMI were calculated using only adults (≥ 20 years old). The Male (Ref.) and Female (Ref.) data are available at e-Stat [The National Statistics Center, 2021a].

Data analysis

The preprocessed data under the latest model schema includes a total of 1868 attributes that came from the 16 measurement categories. This number is comparable to a large cross-sectional study [Nakaji et al., 2021]. Attribute number, missing rate, and outlier rate of each measurement category are shown in S4 Table. Except for Hair Loss Determination and Lipid Mediator Detection, which were relatively difficult to measure, the missing data rate in the other measurement categories was less than 20%, indicating a sufficiently low rate of missing data. Some measurement categories exhibited higher outlier rates than others. Most of these abnormal values were thought to be caused by diseases such as diabetes rather than by noise or other factors, and therefore the data quality was considered to be high. Figure 5 shows the correlation matrix for Data Source A. Many strong relationships existed between variables in the same measurement category, but there were some weak relationships between variables in different measurement categories. This dataset had a sparse data structure, as many pairs of variables had no or weak relationships (see S1 Fig.). Many strong relationships between pairs of variables were observed not only within the same

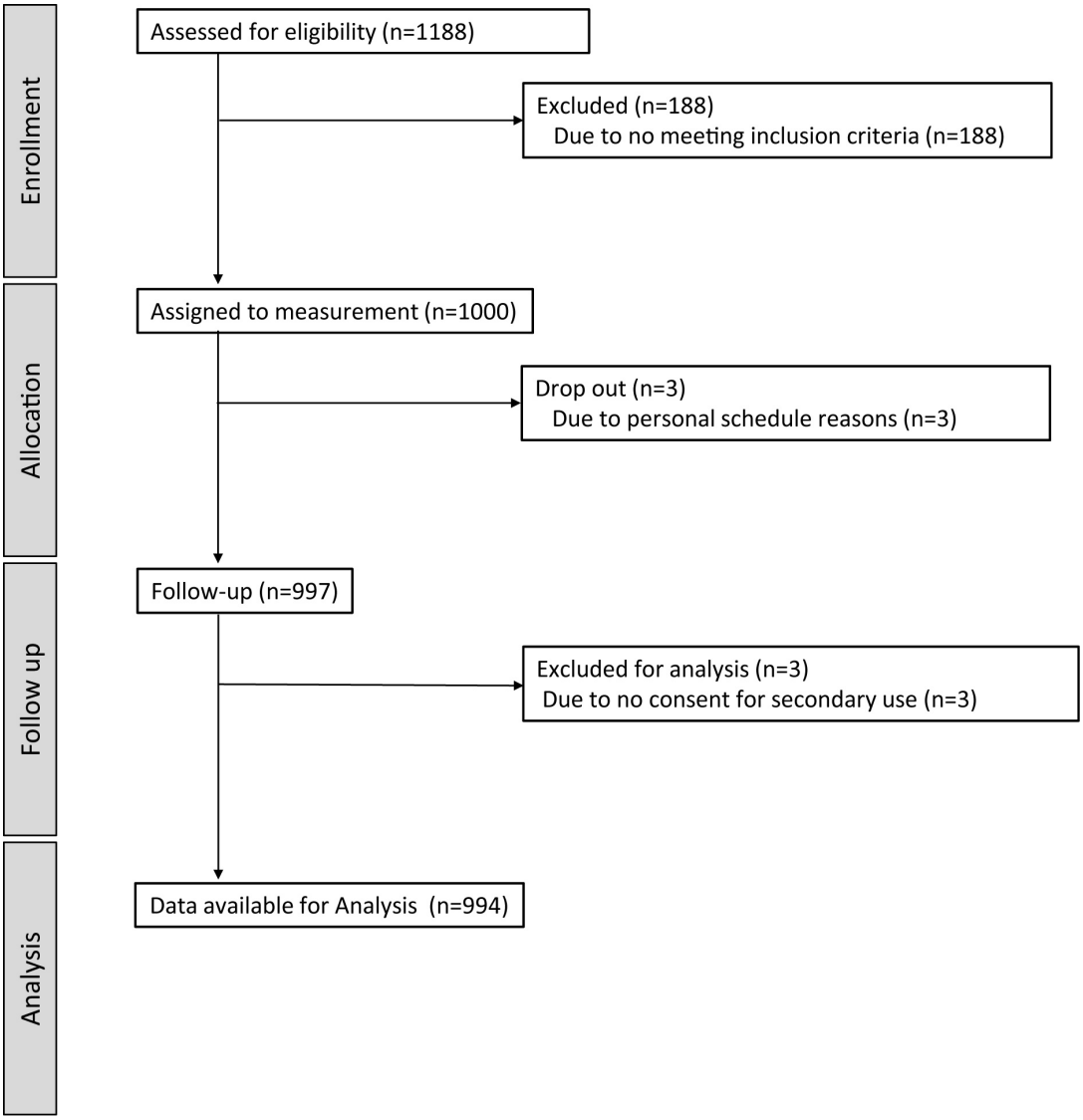

Figure 4. The participants' flow of Data Source A. The number of participants in the entire recruitment process.

measurement category but also across different measurement categories (see several examples in S2 Fig.).

Data Source B

The record and attribute numbers of the preprocessed data were 1,245,807 and 261, respectively, under the latest model schema. The record number corresponds to approximately 1% of adults living in Japan. This data is expected to adequately reflect the statistical characteristics of the annual health checkup and the medical and dental receipts, although the older adult population (≥ 60) was lower due to their retirement and the ratio of males to females was slightly higher due to several possible reasons (e.g., employment, income, lifestyle, *etc.*) [Noguchi and Shen, 2019]. The missing rates were considerably low. Almost all these outliers were thought to be caused by diseases, and thus the data quality was considered to be sufficient.

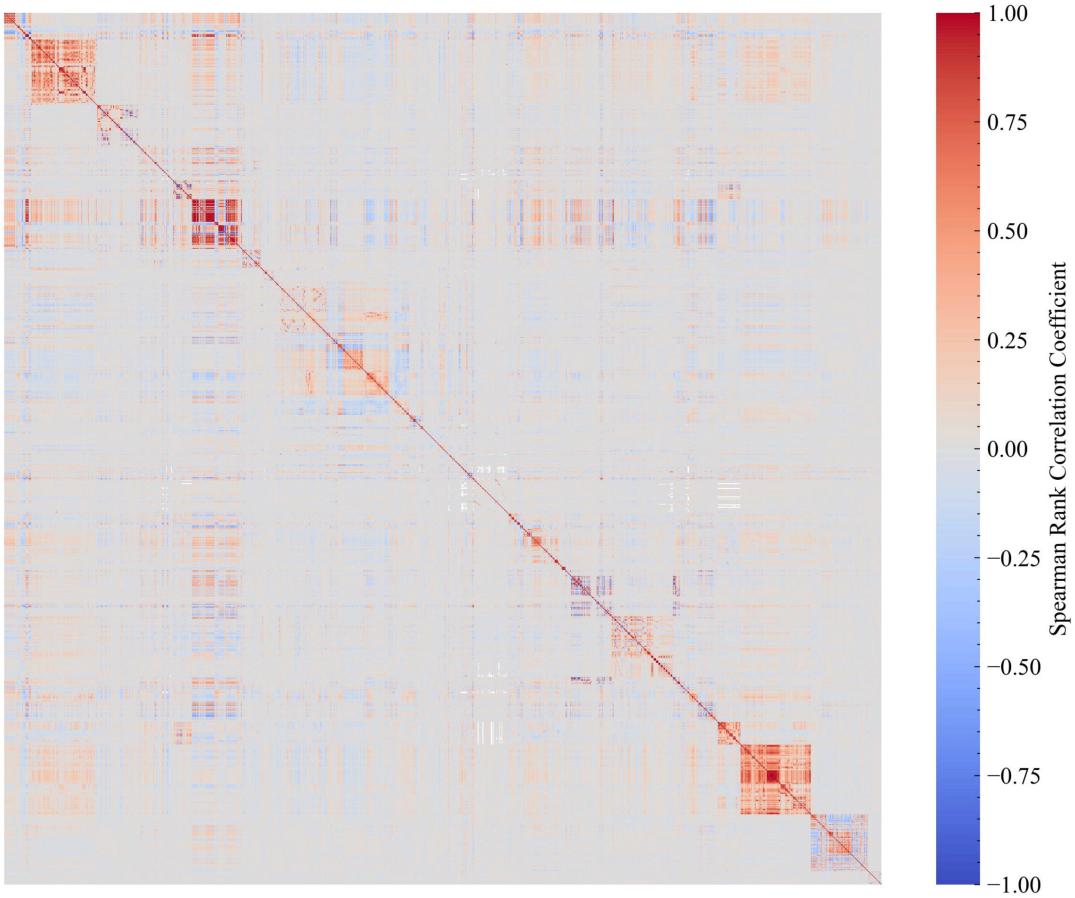

Figure 5. Correlation matrix using the Spearman rank correlation. The color represents the correlation coefficient between real, positive, ordered categorical, and binary categorical attributes.

Data Source C

The record and attribute numbers of the data of Data Source C-1 were 11,646 and 61, respectively, under the latest model schema. The missing rate was considerately low due to the high quality control of the study. Because the study was conducted mainly at workplaces, the ratio of males to females was high. The record and attribute numbers of the aggregated data of Data Source C-2 were 1745 and 162, respectively. The distributions of age, sex, and BMI were not perfectly matched with the overall population in Japan due to the inclusion and exclusion criteria of the intervention studies.

Model schema and training dataset

Table 2 shows the numbers of the preprocessed dataset records. The original datasets were split into 90% for the training dataset and 10% for the holdout dataset; the training dataset was then selected and augmented to incorporate four imbalanced records from various data sources, as illustrated in Table 2. The attribute overlaps from the Data Sources A, B, C-1, and C-2 are described in Fig. 6. Table 3 shows attribute occurrence percentages in Data Sources A, B, C-1, and C-2. As shown in Table 3, common attributes (age, sex, weight, height, BMI, *etc.*) were used to connect all the Data Sources. The number of each type of attribute for each of the data sources is provided in Table 4. As clearly shown in Table 5, this dataset encompasses a range of field categories, defined by the authors based on application fields, from “Vital signs" to “Lifestyle" enabled by multiple data sources. This diversity is attributed to the multiple data sources, particularly Data Source A, which includes various health attributes. Not only “Demographic", but also “General blood testing"

functioned as “common" attributes.

Table 2. Numbers of records of each data source.

Data Source	Original records	90% Records	Adjusted records
A	994	897	18,000 (up sampling)
B	1,245,807	1,121,227	100,000 (down sampling)
C-1	11,646	10,483	18,000 (up sampling)
C-2	1,745	1,584	18,000 (down sampling)

This table shows the number of records in Data Sources A, B, C-1, and C-2. *90% Records* corresponds to the dataset without the holdout set. *Adjusted records* corresponds to the dataset that was selected and augmented for training the model.

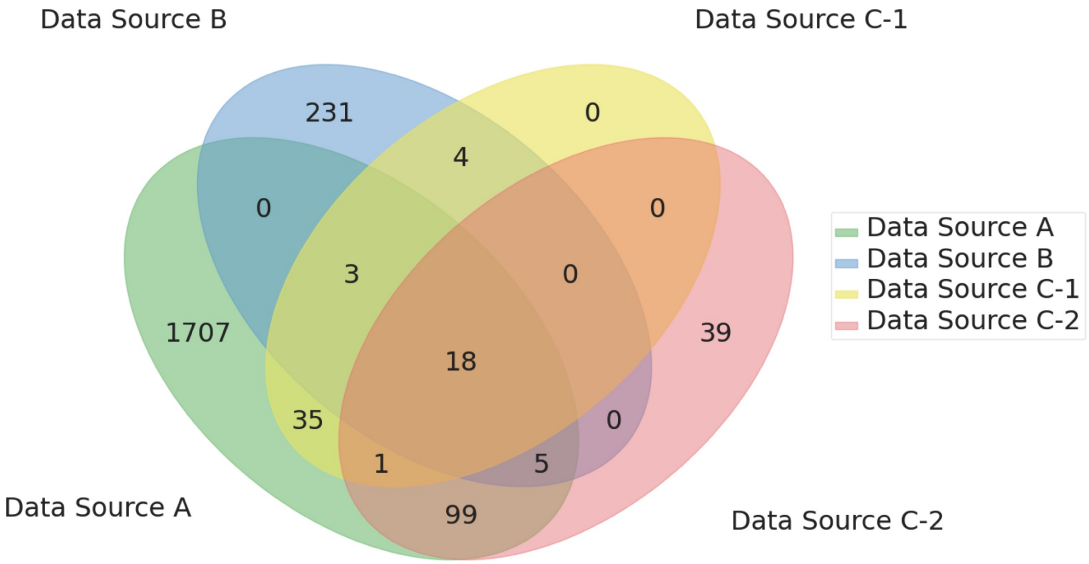

Figure 6. Overlap of attributes across datasets. This image shows attribute overlaps across the Data Sources A, B, C-1, and C-2.

Table 3. Attribute occurrence in all Data Sources.

Occurrence	Number (%)
1	1,977 (92.2)
2	138 (6.4)
3	9 (0.4)
4	18 (0.8)

This table shows the attribute occurrence percentages in Data Sources A, B, C-1, and C-2.

VHGM quality

Missing value imputation

The imputation performance was evaluated using the holdout set, which corresponds to 10% of the original datasets. Since the model output for each attribute is a distribution, we calculated the errors by treating the *mode* of the estimated distribution as if it is the point estimation. The

Table 4. Number of attributes of each data type for the data sources.

Attribute Type	Data Source A	Data Source B	Data Source C-1	Data Source C-2
Real	1,134	15	11	145
Positive	15	10	6	7
Count	0	199	0	0
Categorical	485	27	5	2
Ordered Categorical	234	10	39	8
Total	1,868	261	61	162

The data type definitions are described in the section *Design of model schema*.

means (standard deviations) of standardized errors for real, positive, and count-type attributes were 0.527 (0.550), 0.566 (0.545), and 0.129 (0.589), respectively (Table 6). The means of accuracies for ordered categorical type attributes were lower than those for categorical type attributes (Table 7). This discrepancy is likely due to the greater number of selections for ordered categorical type attributes compared to categorical type attributes. These errors were better than those obtained using the mode imputation for the categorical type, the ordered categorical type, and the count type and the mean imputation for the real type and the positive type, following the same trend observed with the training dataset [Oono et al., 2023]. Thus, this method is practically acceptable for applications involving missing value imputations in setting comparable to our data.

Univariate and bivariate analyses

Figure 7 shows the result examples of univariate and bivariate analyses. With these metrics, we ensured that the model captured important statistical properties of the training dataset. Note that in Fig. 7(b), non-linear relationships between attributes are properly captured, which would not be possible with a simple linear model such as the covariance matrix.

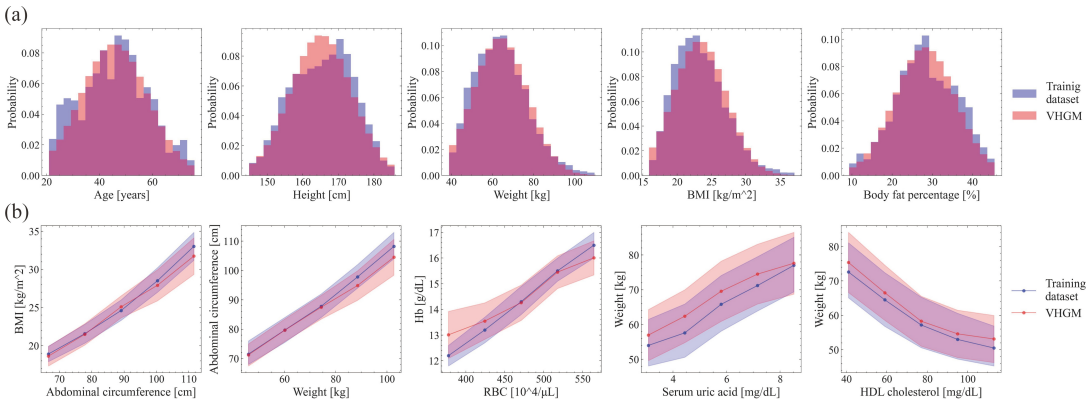

Figure 7. Univariate and bivariate analyses. (a) Univariate comparison between the training dataset and model. (b) Bivariate comparison between the training dataset and model. The blue in the graphs shows the distribution of the training dataset and the red shows the distribution of the model.

Scenario-based analysis

Figure 8 shows the inference results of nutrition intakes with three different lifestyles as an example. Increased intake of carbohydrates was observed in association with unhealthy habits, while the intake of dietary fiber and vitamin C decreased under the same conditions. These results were generally consistent with prior knowledge. Thus, the scenario-based analysis provided generally consistent results. In cases where results contradicted intuitive understanding, we verified the

Table 5. Field categories of the attributes in each data source.

Field	Source A	Source B	Source C-1	Source C-2
Vascular function	53	n/a	n/a	4
Nutrient	243	n/a	n/a	63
Cognitive function	69	n/a	n/a	n/a
Liver function	16	3	3	11
Renal function	17	3	1	14
General blood testing	20	7	4	18
Sugar metabolism function	61	7	3	10
Hormone	10	n/a	n/a	n/a
Stress / Fatigue	49	n/a	n/a	n/a
Body composition / Physique	77	7	5	15
Vital signs	4	2	2	6
Women's health	213	n/a	n/a	1
Skin care / Hair care	193	n/a	n/a	n/a
Motor function	20	2	n/a	n/a
Immunity / Hygiene habit	108	n/a	n/a	n/a
Lifestyle	57	10	38	n/a
Sleep	25	n/a	n/a	n/a
Constitutional classification / Personality	9	n/a	n/a	n/a
Body odor	28	n/a	n/a	n/a
Excretion function	27	n/a	n/a	n/a
Health awareness	76	n/a	n/a	n/a
Oral cavity	33	1	n/a	n/a
Productivity / Presenteeism	36	n/a	n/a	n/a
Demographic	4	2	2	2
Medical history / Medication (questionnaire, number of receipts issued)	55	189	3	n/a
Nursing care	1	n/a	n/a	n/a
Biomarker	181	n/a	n/a	n/a
Walking characteristics	183	1	n/a	1
Medical procedure (number of receipts issued)	n/a	27	n/a	n/a
Other fields	n/a	n/a	n/a	17
Total	1,868	261	61	162

The field categories were defined by the authors, and each attribute was grouped into a single field category in the same way as for the measurement categories in Data Source A.

Table 6. Imputation errors in real, positive, and count-type attributes.

	Real	Positive	Count
Attribute number	1,170	18	199
Absolute standardized error (Mean)	0.527	0.566	0.129
Absolute standardized error (Standard deviation)	0.550	0.545	0.589

The details are provided in S2 Text.

model training algorithm for potential issues and used the findings to enhance the model's performance.

Table 7. Imputation errors in categorical and ordered categorical attributes.

	Categorical	Ordered categorical
Attribute number	511	244
Averaged accuracy (Mean)	0.839	0.573

The details are provided in S2 Text.

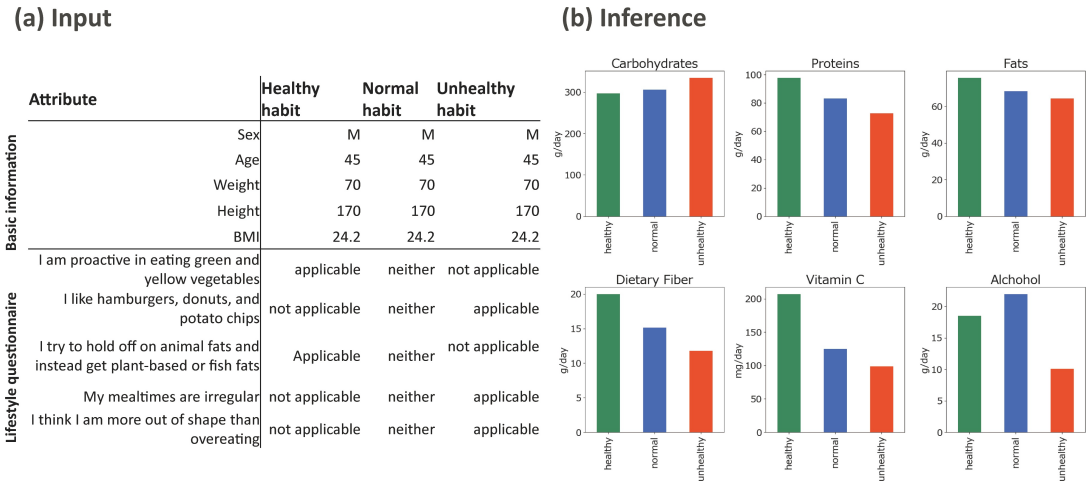

Figure 8. Scenario-based analysis. (a) Anticipated inputs using the same basic information with three different lifestyles: healthy habits, normal habits, and unhealthy habits. (b) Inference of nutrient intake per day with the three different lifestyles.

External dataset validation

The results of the external validation are summarized in Table 8 and Table 9 , which detail the z-scores and overlap areas from the comparison between the VHGM output and the e-Stat and NHANES datasets, respectively. The analysis indicates a higher degree of statistical consistency between the VHGM output and the e-Stat data (mean overlap area = 0.82) compared to the NHANES data (mean overlap area = 0.59). This finding is expected, as the model's training data was sourced from a Japanese population, which is more demographically similar to the population represented in e-Stat than to the US-based population in NHANES.

Table 8. Overview of the statistical difference between VHGM and e-Stat.

Metric	Mean	Std	Max	Min
Absolute z-score	0.15	0.08	0.28	0.01
Overlap area	0.82	0.05	0.90	0.70

The table shows summary statistics for the 24 common attributes. The absolute z-score was calculated by treating the mean from the VHGM output as a sample mean relative to the distribution of the corresponding attribute in the e-Stat dataset. The overlap area quantifies the similarity between the two distributions.

Comparison with other generative models

Table 10 shows the results of a regression task for TVAE, CTGAN [Xu et al., 2019], and Gaussian Copula along with our algorithm [Oono et al., 2023]. Our algorithm outperforms the known generative models for the synthetic data (the "Individual" column in the table). Interestingly, we can obtain bet-

Table 9. Overview of the statistical difference between VHGM and NHANES.

Metric	Mean	Std	Max	Min
Absolute z-score	0.38	0.23	0.82	0.01
Overlap area	0.59	0.15	0.84	0.35

Absolute z-scores and overlap areas were evaluated using the same methodology as in Table 8, with NHANES as the external comparison dataset.

ter performance if the synthetic data is used for augmenting the original data (the second column in the table).

Table 10. Comparison with other algorithms on a regression task.

Dataset	Individual (R^2)	Combined with original (R^2)
Original	0.531	–
TVAE	0.225	0.525
CTGAN	-0.130	0.506
Gaussian Copula	0.476	0.531
Ours	**0.501**	**0.553**

Machine Learning Efficiency: the task is to estimate from 65 attributes related to exercise, nutrients, sleeping habits, stress, and tiredness. We generated 10,000 synthetic records for the 66 attributes. R^2 denotes the coefficient of determination.

Governance process

Table 11 summarizes the models published as of August 2025, following the decisions of the multi-stakeholder committee. The number of attributes has generally increased with addition of new clinical data into the data sources and introduction of new attributes in the model schema. Model inference performance has also improved due to incremental algorithm updates [Oono et al., 2023]. As shown in Fig. 9, model inference performance in bivariate analysis has improved.

Although the training dataset of the VHGM does not contain any personally identifiable information, the improvement of model inference performance may increase potential privacy concerns regarding *membership inference attacks*, which aim to determine if a particular individual was included in the training dataset. As recommended by the committee, we conducted a preliminary assessment vulnerability to membership inference attacks and reviewed the results. A summary of these preliminary experiments has been reported elsewhere [Bito et al., 2023], indicating no immediate evidence of exploitable risk. Nevertheless, we acknowledge that a more comprehensive empirical evaluation is warranted and plan to address this in future work.

Discussion

Principal findings

The principal findings of this paper are twofold. First, a general-purpose generative model for healthcare can be built by combining multiple data sources and managing its quality to a certain degree. Second, there are identifiable *patterns* for how such a model can be applied in real-world applications.

Feasibility of a general-purpose statistical model

The VHGM is provided as an API and is intended to be a building block for healthcare applications developed and operated by independent vendors. We demonstrate two such applications.

Table 11. Published models as an API service and their model parameter sizes.

Model name	Date of publish	Feature	Attributes	Parameter size
fomalhaut	Dec. 7, 2022	General-purpose	1,841	1,668,232
helvetios-a	Feb. 2, 2023	Health checkup focus	260	290,266,849
helvetios-b	Feb. 2, 2023	Health checkup focus	302	11,056,686
intercrus	May 12, 2023	General-purpose	1,975	85,822,679
libertas	June 30, 2023	General-purpose	2,150	94,131,094
lich	Oct. 6, 2023	General-purpose	2,097	26,951,129
musica	Apr. 8, 2024	General-purpose	2,110	32,381,324
pollux	May 21, 2025	General-purpose	2,142	2,736,600

The model parameter size is typically dominated by the number of hidden nodes and the number of layers in the model architecture. The model names are based on the names of stars.

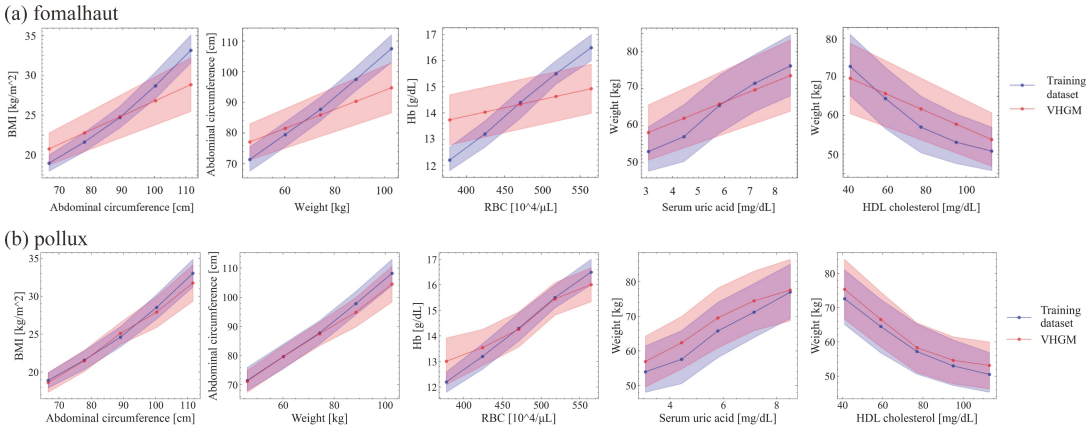

Figure 9. Improved model performance in bivariate relationships. Overlap areas between the dataset and model output were increased for each attribute pair. (a) Bivariate relationships using *fomalhaut* published in Oct. 2022. (b) Bivariate relationships using *pollux* published in May 2025.

1. **App for Encouraging More Walking:** A mobile phone company has developed a healthcare app for their phones, which encourages users to walk more for their health. One of the challenges for the app is setting appropriate goals (the number of steps a user should walk daily) because different people have different abilities. The VHGM includes attributes such as daily walking steps and other factors like lower back pain. The app uses this information to suggest, “People like you but without back pain walk this number of steps daily on average," allowing users to decide whether to try walking more.
2. **App for Health-Related Financial Assets:** The VHGM contains attributes derived from health insurance records. With these attributes, one can estimate the distribution of annual medical spending, given available attribute values such as age, sex, weight, and lifestyle habits. A financial service startup uses this information to calculate the estimated lifetime medical spending, referred to as “Health Asset." The app enables users to know their Health Asset and experiment with how this number changes based on different habits, such as exercise, drinking, and smoking. To prevent misunderstanding or misuse, the service clearly states that this application is not a medical device and is not intended for the diagnosis, treatment, or prevention of any disease. Users are advised to interpret the results as informative guidance rather than definitive medical or financial advice.

These applications provide evidence of the usefulness of the VHGM as a general-purpose generative model. More applications of the VHGM are described in the model development paper [Oono

et al., 2023].

Usage patterns of the VHGM

As of the time of writing this paper, there are several paying customers who regularly use the VHGM. Additionally, we conducted a couple of business idea contests, asking for new applications based on the VHGM. From these experiences, we observed certain patterns in how the VHGM can be used.

1. **Estimation of a missing value from known values** – This is the basic function of the VHGM. Given observed values $o_1, o_2, ..., o_m$, VHGM returns the estimated distribution $P(y|o_1, o_2, ..., o_m)$ for the target attribute y. This pattern is useful when some attribute is difficult to measure directly (e.g., measuring blood sugar usually requires an invasive process – VHGM provides a means to estimate the blood sugar from other observable attributes).
2. **What-if analysis** (Counter-factual scenario generation) – One can provide counter-factual input to the VHGM. For example, “what would my estimated BMI be if I were not smoking" is a counter-factual query. This should not be interpreted as a causality, but these queries are useful for considering possible scenarios and planning future course of action. The walking encouragement app described above uses this pattern.
3. **Optimization for a desired output** – One can use the VHGM API to iteratively search possible combinations of values that would yield the desired estimated value of the output attribute. For example, “How can I change my diet to reduce the estimated risk of neuropathic pain" would be answered by optimizing the diet attributes to reduce the estimated number of annual doctor visits related to neuropathic pain. Open-source tools for black-box optimization such as Optuna [Akiba et al., 2019] could be used for such computations.
4. **Exploration of possible factors** – One can explore possible attributes that have some relationship with the target attribute. For example, many older adults are concerned about their body odor but are unaware of the factors that may influence it. By changing the value of the body odor attributes and observing how the other over 2000 attributes respond, one may be able to form a hypothesis on the cause of body odor.

For each of the above examples, there are two “modes" of using the VHGM. One is to use it through the API to directly obtain the query results. The other is to generate synthetic data under given conditions and then use the synthetic data for further analysis. In general, this “indirect" mode of use is not recommended because the resulting analysis may contain both the errors incurred by the VHGM training process and the errors in the second, derivational analysis. This mode is, however, useful for:

- **Educational purposes** because the trainees do not need to have a programming environment for API access
- **Analysis in specific groups** for which no data is available.

This is by no means an exhaustive list. We expect that there will be more innovative VHGM use cases in the future.

Limitations

Limitations of the VHGM are as follows:

- **Bias in the training dataset** – Due to the cost, logistical constraints, and clinical study purposes, the populations used for building the training dataset were biased. This may make applications that target different populations (e.g., different racial groups) inappropriate. The external dataset validation indicates this limitation. Certain medical conditions specific to populations or environments outside the scope of the training data may not be appropriately addressed.

- **Comprehensiveness of evaluation** – While we devised a multi-faceted evaluation framework (imputation error, distributional concordance, and scenario-based analysis), it does not comprehensively capture all aspects of model quality. Moreover, due to the general-purpose design of the VHGM, it is not feasible to exhaustively benchmark all possible input-output combinations. This limited comprehensiveness should be considered when interpreting the reported results.
- **Deep stratified analysis** – Due to the available sizes of the source datasets, including that of Data Source A, deeply stratified groups do not have enough records, which may result in an unreliable statistical model.
- **Cross-sectionality** – The training dataset is largely cross-sectional (Data Source C-2 was from interventional studies) and therefore, the VHGM is not capable of predicting the future. It is theoretically possible to build a model with predictive functions if we have high-dimensional time-series data, but collecting such data is excessively expensive.
- **Correlation vs causality** – *What-if analysis* carries the risk of being interpreted as causal. If the VHGM returns "if you were doing daily exercise, your estimated BMI would be lower," it does not mean that exercise will lower the BMI. Clear communication of how to interpret the output of the VHGM is one of the critical risk factors we identified. Transparency (see Governance process) is one of the mitigation efforts.

Conclusion

We demonstrated the feasibility of constructing a general-purpose generative model for healthcare data. Our analysis confirms that the model captures key statistical properties, including univariate distributions and bivariate relationships among attributes. Additionally, we presented several real-world applications to highlight the model's practical value.

Acknowledgments

The primary sponsor of this study is Kao Corporation (Tokyo, Japan). None of the authors have received any specific funding for this study. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. KB, Y Saito, MS, SK, and AK are employees of Kao Corporation (Tokyo, Japan). KO, NC, ZG, HI, MY, YO, HO, KA, SY, Y Sugawara, and SM are employees of Preferred Networks Incorporated (Tokyo, Japan). HM is an executive fellow of Kao Corporation and a senior advisor of Preferred Networks Incorporated.

The authors thank Koki Tsuda, Satoru Mochizuki, Seoyoun Chung, Takahito Nakamura, Juntaro Yamashita, Tetsuya Uchiyama, Kunihiko Miyoshi, Masahiro Hirasawa, Tsukasa Takemura, Takahiro Takamuku, Tetsuya Yamaguchi, Aiko Suzuki, Shota Hayashi, and Kohei Hayashi for their support for the development of the models and applications. We also thank Kei Sugitani, Adeline Muliandi, Nami Yamanaka, Takahiro Hasumura, Yasutoshi Ando, Takashi Fushimi, Teruhisa Fujimatsu, Tomoki Akatsu, Sawako Kawano, Ren Kimura, Shigeki Tsuchiya, Yuuki Yamamoto, Mai Haneoka, Ken Kushida, Tomoki Hideshima, Eri Shimizu, Jumpei Suzuki, Aya Kirino, Hisashi Tsujimura, Shun Nakamura, Takashi Sakamoto, Yuki Tazoe, Masayuki Yabuki, Shinobu Nagase, Tamaki Hirano, Reiko Fukuda, Yukari Yamashiro, Yoshinao Nagashima, Nobutoshi Ojima, Motoki Sudo, Naoki Oya, Yoshihiko Minegishi, and Koichi Misawa for their analysis to identify the numerous health attributes across diverse categories.

We acknowledge the valuable contributions of Professor Kazuhiro Minami at The Institute of Statistical Mathematics (Tokyo, Japan) through insightful discussions on privacy issues.

We are deeply grateful to MinaCare Co., Ltd. and its founder, Yuji Yamamoto, for providing the commercial healthcare dataset under flexible terms and conditions. Without their belief in the positive impact of widespread data dissemination on healthcare, this project would not have been realized.

Supplementary

S1 Fig.

Histogram of correlation coefficients using the Spearman rank correlation. The number of combinations of selecting pairs of attributes from 1776 attributes is 1,088,550. For each combination, the Pearson rank correlation coefficient was obtained, and the histogram was created.

S2 Fig.

Typical correlated pairs of each data type in Data Source A. The Φ_K correlation coefficient (Baak et al., 2020) was employed to assess the relationships between pairs of variables across various data types, including numerical, categorical, and ordinal. Each 5-letter code corresponds to an attribute definition. (a) Typical correlated pairs between numerical attributes. The numbers in the titles represent the correlation coefficient of the pairs. (b) Typical correlated pairs between a numerical attribute and a categorical attribute. The numbers in the titles represent the correlation coefficient of the pairs. (c) Typical correlated pairs between categorical attributes. The numbers in the titles represent the correlation coefficient of the pairs.

S1 Text.

Analytical method details on skin surface lipid (SSL)-RNA, intestinal microbiota, and saliva microbiota. Method details are described in S1 Text.

S2 Text.

Details on the imputation error calculation methods. Method details are described in S2 Text.

S3 Text.

Details on the univariate and bivariate analyses. Method details are described in S3 Text.

S4 Text.

Details on the external dataset validation. Method details are described in S4 Text.

S1 Table.

Number of male participants. Decade, N (visit 1), Ratio (%), N (visit 2), Conversion (%), and Ref. (%) correspond to the age group of participants, numbers in each age group, percentages of each age group relative to the total participants, ratios of N (visit 2) to N (visit 1), and ratios of the Japanese population from the recent Japanese official statistics that are available at e-Stat (2019), respectively. At visit 2, conversion rates of the men's 60-69 and $\geq$ 70 age groups were slightly lower compared to those in younger age groups. This reduction appears to be attributable to the exclusion criteria applied in the study or their health conditions (see the medical history records in S3 Table).

S2 Table.

Number of female participants. The column name definitions are the same as S1 Table.

S3 Table.

Number of diseases under treatment. Numbers of participants undergoing treatment for each disease.

S4 Table.

Attribute number, missing rate, and outlier rate of each measurement. Measurement categories were described elsewhere [Hibi et al., 2023].

References

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama (2019). "Optuna: A next-generation hyperparameter optimization framework". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631.

Alhussein, Musaed and Ghulam Muhammad (2018). "Voice pathology detection using deep learning on mobile healthcare framework". In: *IEEE Access* 6, pp. 41034–41041.

Artico, Fausto, Arthur L Edge III, and Kyle Langham (2022). "The future of artificial intelligence for the BioTech big data landscape". In: *Current Opinion in Biotechnology* 76, p. 102714.

Baak, Max, Rose Koopman, Hella Snoek, and Sander Klous (2020). "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics". In: *Computational Statistics & Data Analysis* 152, p. 107043.

Bito, K., M. Hibi, K. Oono, Y. Saito, K. Minami, and H. Maruyama (2023). "Quality management on Virtual Generative Human Model". In: *Proceedings of the 40th Annual Meeting of the Japan Society for Software Science and Technology*. in Japanese, 53–R.

Centers for Disease Control and Prevention, National Center for Health Statistics (2021). *NHANES Questionnaires, Datasets, and Related Documentation (2017–2020). National Health and Nutrition Examination Survey*. URL: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020.

Chikama, A, T Yamaguchi, T Watanabe, K Mori, Y Katsuragi, I Tokimitsu, O Kajimoto, and M Kitakaze (2006). "Effects of chlorogenic acids in hydroxyhydroquinone-reduced coffee on blood pressure and vascular endothelial function in humans". In: *Prog Med* 26, pp. 1723–1736.

Čolaković, Alem and Mesud Hadžialić (2018). "Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues". In: *Computer networks* 144, pp. 17–39.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv preprint arXiv:2010.11929*.

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (2022). "Masked Autoencoders Are Scalable Vision Learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hibi, Masanobu, Shun Katada, Aya Kawakami, Kotatsu Bito, Mayumi Ohtsuka, Kei Sugitani, Adeline Muliandi, Nami Yamanaka, Takahiro Hasumura, Yasutoshi Ando, et al. (2023). "Assessment of Multidimensional Health Care Parameters Among Adults in Japan for Developing a Virtual Human Generative Model: Protocol for a Cross-sectional Study". In: *JMIR Research Protocols* 12.1, e47024.

Hsiao, Wesley Wei-Wen, Jui-Chu Lin, Chien-Te Fan, and Saint Shiou-Sheng Chen (2022). "Precision health in Taiwan: A data-driven diagnostic platform for the future of disease prevention". In: *Computational and Structural Biotechnology Journal* 20, pp. 1593–1602.

Kao Corporation (2023). *News Letter, VITA-GATE*. Accessed: 2025-01-08. URL: https://vita-gate.com/newsletters.

Kozuma, K, A Chikama, E Hoshino, K Kataoka, K Mori, T Hase, Y Katsuragi, I Tokimitsu, and H Nakamura (2005). "Effect of intake of a beverage containing 540 mg catechins on the body composition of obese women and men". In: *Prog Med* 25.7, pp. 1945–57.

Kreuzberger, Dominik, Niklas Kühl, and Sebastian Hirschl (2023). "Machine learning operations (mlops): Overview, definition, and architecture". In: *IEEE access* 11, pp. 31866–31879.

Lu, Zhao-xia, Peng Qian, Dan Bi, Zhe-wei Ye, Xuan He, Yu-hong Zhao, Lei Su, Si-liang Li, and Zheng-long Zhu (2021). "Application of AI and IoT in clinical medicine: summary and challenges". In: *Current medical science* 41.6, pp. 1134–1150.

Maruyama, Hiroshi (2018). "Guaranteeing Deep Neural Network Outputs in a Feasible Region". In: *Proceedings of the International Workshop on Evidence-based Security and Privacy in the Wild and the 1st International Workshop on Machine Learning Systems Engineering*. CEUR Workshop Proceedings, p. 3.

Matsui, Y, K Kinoshita, N Osaki, T Wakisaka, M Hibi, Y Katsuragi, TF Yamaguchi, and I Fukuhara (2018). "Effects of Tea Catechin-rich Beverage on Abdominal Fat Area and Body Weight in Obese

Japanese Individuals - A Randomized, Double-blind, Placebo-controlled, Parallel-group Study -". In: *Jpn Pharmacol Ther* 46.8, pp. 1383–1395.

Matsui, Y, M Takeshita, M Hibi, I Fukuhara, and N. Osaki (2016). "Efficacy and safety of powdered beverage containing green tea catechins on body fat in obese adults - a randomized, placebo controlled, double-blind parallel study". In: *Jpn Pharmacol Ther* 44.7, pp. 1013–1023.

Ministry of Justice, Japan (2017). *Act on Anonymized Medical Data That Are Meant to Contribute to Research and Development in the Medical Field (Act No. 28 of 2017), Japanese Law Translation Database System*. URL: https://www.japaneselawtranslation.go.jp/en/laws/view/3441.

Nagao, T, R Ochiai, Y Katsuragi, Y Hayakawa, K Kataoka, M Komikado, I Tokimitsu, and T Tsuchida (2007). "Hydroxyhydroquinone-reduced milk coffee decreases blood pressure in individuals with mild hypertension and high-normal blood pressure". In: *Prog Med* 27, pp. 2649–2664.

Nagao, Tomonori, Tadashi Hase, and Ichiro Tokimitsu (2007). "A green tea extract high in catechins reduces body fat and cardiovascular risks in humans". In: *Obesity* 15.6, pp. 1473–1483.

Nagao, Tomonori, Ryuji Ochiai, Takuya Watanabe, Kiyoshi Kataoka, Masanori Komikado, Ichiro Tokimitsu, and T Tsuchida (2009). "Visceral fat–reducing effect of continuous coffee beverage consumption in obese subjects". In: *Jpn Pharmacol Ther* 37.4, pp. 333–344.

Nakaji, Shigeyuki, Kazushige Ihara, Kaori Sawada, Stefano Parodi, Takashi Umeda, Ippei Takahashi, Koichi Murashita, Shizuka Kurauchi, and Itoyo Tokuda (2021). "Social innovation for life expectancy extension utilizing a platform-centered system used in the Iwaki health promotion project: A protocol paper". In: *SAGE Open Medicine* 9, p. 20503121211002606.

Noguchi, Riko and Junyi Shen (2019). "Factors affecting participation in health checkups: Evidence from Japanese survey data". In: *Health Policy* 123.4, pp. 360–366.

Oono, Kenta, Nontawat Charoenphakdee, Kotatsu Bito, Zhengyan Gao, Yoshiaki Ota, Shoichiro Yamaguchi, Yohei Sugawara, Shin-ichi Maeda, Kunihiko Miyoshi, Yuki Saito, Koki Tsuda, Hiroshi Maruyama, and Kohei Hayashi (2023). *Virtual Human Generative Model: Masked Modeling Approach for Learning Human Characteristics*. arXiv: 2306.10656 [cs.LG]. URL: https://arxiv.org/abs/2306.10656.

Pathak, Aditi, Laina Serrer, Daniela Zapata, Raymond King, Lisa B Mirel, Thomas Sukalac, Arunkumar Srinivasan, Patrick Baier, Meera Bhalla, Corinne David-Ferdon, et al. (2024). "Privacy preserving record linkage for public health action: opportunities and challenges". In: *Journal of the American Medical Informatics Association* 31.11, pp. 2605–2612.

Pezoulas, Vasileios C, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis (2024). "Synthetic data generation methods in healthcare: A review on open-source tools and methods". In: *Computational and Structural Biotechnology Journal*.

Pradhan, Bikash, Saugat Bhattacharyya, and Kunal Pal (2021). "IoT-based applications in healthcare devices". In: *Journal of healthcare engineering* 2021.1, p. 6632599.

Takase, Hideto, Tomonori Nagao, Kazuhiro Otsuka, Kazuya Kozuma, Shinichi Meguro, Masanori Komikado, and Ichiro Tokimitsu (2008). "Effects of long-term ingestion of tea catechins on visceral fat accumulation and metabolic syndrome – Pooling analysis of 7 randomized controlled trials". In: *Jpn Pharmacol Ther* 36.6, pp. 509–514.

Takase, Hideto, Naoki Sakane, Toshihisa Morimoto, Takanobu Uchida, Kenta Mori, Mitsuhiro Katashima, and Yoshihisa Katsuragi (2019). "Development of a Dietary Factor Assessment Tool for Evaluating Associations between Visceral Fat Accumulation and Major Nutrients in Japanese Adults". In: *Journal of Obesity* 2019, p. 9497861. ISSN: 2090-0708. DOI: 10.1155/2019/9497861. URL: https://doi.org/10.1155/2019/9497861.

Takeshita, Masao, Shinichiro Takashima, Ushio Harada, Eiichiro Shibata, Naoki Hosoya, Hideto Takase, Kazuhiro Otsuka, Shinichi Meguro, Masanori Komikado, and Ichiro Tokimitsu (2008). "Effects of long-term consumption of tea catechins-enriched beverage with no caffeine on body composition in humans". In: *Jpn Pharmacol Ther* 36, pp. 767–776.

The National Statistics Center (2021a). *Average height and weight and standard deviations, National Health and Nutrition Survey*. Accessed: 2025-01-08. URL: https://www.e-stat.go.jp/dbview?sid=0003224177.

— (2021b). *Nutrients Intake, National Health and Nutrition Survey*. Accessed: 2025-08-27. URL: https://www.e-stat.go.jp/dbview?sid=0003234480.

Tsuchida, T, H Itakura, and H. Nakamura (2002). "Reduction of body fat in humans by long-term ingestion of catechins". In: *Prog Med* 22, pp. 2189–2203.

Von Elm, Erik, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, Jan P Vandenbroucke, Strobe Initiative, et al. (2014). "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies". In: *International journal of surgery* 12.12, pp. 1495–1499.

Wang, Wei, Yuxiang Yan, Zheng Guo, Haifeng Hou, Monique Garcia, Xuerui Tan, Enoch Odame Anto, Gehendra Mahara, Yulu Zheng, Bo Li, et al. (2021). "All around suboptimal health—a joint position paper of the Suboptimal Health Study Consortium and European Association for Predictive, Preventive and Personalised Medicine". In: *EPMA Journal* 12, pp. 403–433.

Watanabe, Takuya, Shinichi Kobayashi, Tohru Yamaguchi, Masanobu Hibi, Ikuo Fukuhara, and Noriko Osaki (2019). "Coffee abundant in chlorogenic acids reduces abdominal fat in overweight adults: A randomized, double-blind, controlled trial". In: *Nutrients* 11.7, p. 1617.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni (2019). "Modeling Tabular Data Using Conditional GAN". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Yamaguchi, T., A. Chikama, M. Inaba, R. Ochiai, Y. Katsuragi, I. Tokimitsu, T. Tsuchida, and I. Saito (2007). "Antihypertensive effects of hydroxyhydroquinone-reduced coffee on high-normal blood pressure". In: *Prog Med* 27, pp. 683–694.