Early Diagnosis of Hereditary Angioedema Based on US Medical Dataset: Algorithm Development and Validation in Japan

Kouhei Yamashita, Yuji Nomoto, Tomoya Hirose, Akira Yutani, Akira Okada, Nayu Watanabe, Ken Suzuki, Munenori Senzaki, Tomohiro Kuroda

Submitted to: JMIR Medical Informatics
on: April 26, 2024

Table of Contents

Early Diagnosis of Hereditary Angioedema Based on US Medical Dataset: Algorithm Development and Validation in Japan

Kouhei Yamashita[1] MD, PhD; Yuji Nomoto[2] MD; Tomoya Hirose[3] MD, PhD; Akira Yutani[4] PhD; Akira Okada[5] ME; Nayu Watanabe[5] MMG; Ken Suzuki[5] ME; Munenori Senzaki[5] MS; Tomohiro Kuroda[4] PhD

[1]Department of Hematology and Oncology, Graduate School of Medicine Kyoto University Kyoto JP
[2]Department of Palliative Care Medicine Niigata City General Hospital Niigata JP
[3]Department of Traumatology and Acute Critical Medicine Osaka University Graduate School of Medicine Osaka JP
[4]Division of Medical Information Technology and Administration Planning Kyoto University Hospital Kyoto JP
[5]Healthcare and Life Science, IBM Consulting IBM Japan, Ltd Tokyo JP

Corresponding Author:
Kouhei Yamashita MD, PhD
Department of Hematology and Oncology, Graduate School of Medicine
Kyoto University
54 Shogoin-kawahara-cho, Sakyo-ku
Kyoto
JP

Abstract

Background: The rare genetic disease hereditary angioedema (HAE) induces acute attacks of swelling in various regions of the body. Its prevalence is estimated to be 1 in 50,000 people, with no reported bias among different ethnic groups. However, considering the estimated prevalence, the number of patients in Japan diagnosed with HAE remains approximately 1 in 250,000, which means that only 20% of potential HAE cases are identified.

Objective: We aimed to develop an artificial intelligence (AI) model that can detect patients with suspected HAE using medical history data (medical claims, prescriptions, and EMR) in the US and validate the detection performance of HAE cases. We also aimed to verify whether the model was applicable to Japanese data.

Methods: The HAE patient and control groups were identified from the US claims and EMR datasets. We analyzed the characteristics of the diagnostic history of patients with HAE and developed an AI model to predict the probability of HAE based on a generalized linear model and bootstrap method. The model was then applied to the EMR data of the Kyoto University Hospital to verify its applicability in Japanese data.

Results: Precision and sensitivity were measured to validate the model performance. The precision score was 2% in the initial model-development step using the comprehensive US dataset. This means that while the prevalence of HAE is 1/50,000, our model can screen out suspected patients, where 1 in 50 of these patients actually have HAE. In addition, in the validation step with Japanese EMR data, the precision score was 23.6%, which exceeded our expectations. We achieved 61.5% sensitivity in the US and 37.6% in the validation in a single Japanese hospital. Overall, our model predicted patients with typical HAE symptoms.

Conclusions: This study indicates that our AI model can detect HAE in patients with typical symptoms and is effective in Japanese data. However, further prospective clinical studies are required to investigate whether this model can be used to diagnose HAE.

(JMIR Preprints 26/04/2024:59858)
DOI: https://doi.org/10.2196/preprints.59858

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Kouhei Yamashita, MD, PhD;
Department of Hematology and Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan.

Yuji Nomoto, MD;
Department of Palliative Care Internal Medicine, Niigata City General Hospital, Niigata, Japan.

Tomoya Hirose, MD, PhD;
Department of Traumatology and Acute Critical Medicine, Osaka University Graduate School of Medicine, Osaka, Japan.

Akira Yutani, PhD;
Division of Medical Information Technology and Administration Planning, Kyoto University Hospital, Kyoto, Japan.

Akira Okada, ME;
Healthcare and Life Science, IBM Consulting, IBM Japan, Ltd, Tokyo, JP

Nayu Watanabe, MMG;
Healthcare and Life Science, IBM Consulting, IBM Japan, Ltd, Tokyo, JP

Ken Suzuki, ME;
Healthcare and Life Science, IBM Consulting, IBM Japan, Ltd, Tokyo, JP

Munenori Senzaki, MS;
Healthcare and Life Science, IBM Consulting, IBM Japan, Ltd, Tokyo, JP

Tomohiro Kuroda, PhD
Division of Medical Information Technology and Administration Planning, Kyoto University Hospital, Kyoto, Japan.

All authors contributed equally.

Early Diagnosis of Hereditary Angioedema Based on US Medical Dataset: Algorithm Development and Validation in Japan

Abstract

Background:
Hereditary angioedema (HAE), a rare genetic disease, induces acute attacks of swelling in various regions of the body. Its prevalence is estimated to be 1 in 50,000 people, with no reported bias among different ethnic groups. However, considering the estimated prevalence, the number of patients in Japan diagnosed with HAE remains approximately 1 in 250,000, which means that only 20% of potential HAE cases are identified.
Objective:
We aimed to develop an artificial intelligence (AI) model that can detect patients with suspected

HAE using medical history data (medical claims, prescriptions, and EMR) in the US. We also aimed to validate the detection performance of the model for HAE cases using the Japanese dataset.
Methods:
The HAE patient and control groups were identified using the US claims and EMR datasets. We analyzed the characteristics of the diagnostic history of patients with HAE and developed an AI model to predict the probability of HAE based on a generalized linear model and bootstrap method. The model was then applied to the EMR data of the Kyoto University Hospital to verify its applicability for Japanese dataset.
Results:
Precision and sensitivity were measured to validate the model performance. The precision score was 2% in the initial model-development step using the comprehensive US dataset. Our model can screen out suspected patients, where 1 in 50 of these patients have HAE. In addition, in the validation step with Japanese EMR data, the precision score was 23.6%, which exceeded our expectations. We achieved a sensitivity score of 61.5% for the US dataset and that of 37.6% for the validation exercise using data from a single Japanese hospital. Overall, our model could predict patients with typical HAE symptoms.
Conclusions:
This study indicates that our AI model can detect HAE in patients with typical symptoms and is effective in Japanese data. However, further prospective clinical studies are required to investigate whether this model can be used to diagnose HAE.

Keywords:
machine learning; screening, AI; prediction; rare diseases; HAE; electronic medical record (EMR); real world data (RWD); big data

Introduction

The rare genetic disease hereditary angioedema (HAE) induces acute attacks of swelling in various regions of the body, including the face, hands, arms, legs, abdomen, genitals, buttocks, and throat. Gastrointestinal disturbances such as abdominal pain, nausea, and vomiting are frequently associated with edema. Laryngeal edema is rare, even though more than half of the patients with HAE encounter this life-threatening condition [1]. Its global prevalence is estimated to be 1 in 50,000 people, with no reported bias among different ethnic groups [2]. In Japan, about 1 in 250,000 people are diagnosed with HAE, which suggests that only 20% of potential HAE cases are identified [3], suggesting that many patients with HAE remain undiagnosed in Japan. Moreover, in Japan, the mean duration from the first symptoms to diagnosis is 15.6 years [4], which is longer than that in Europe and the United States [5,6]. Early detection of undiagnosed patients is critical for effective treatment of HAE.

To overcome this situation in Japan, the Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema (DISCOVERY) was established in 2021 [7]; it aimed to identify undiagnosed patients with HAE and provide them with appropriate treatment as early as possible.

In this study, we aimed to develop an artificial intelligence (AI) model that can detect suspected HAE patients using medical history data (claims and electronic medical records [EMR]) in the US. We then sought to validate the performance of the model in detecting HAE cases. Additionally, we conducted a pilot study at Kyoto University Hospital (KUHP) using the EMR data to verify the applicability of whether the model was applicable for medical data obtained from the Japanese population. The main objective of this study was to verify whether this model could identify patients

with a history of HAE and/or related diseases.

Methods

First, we developed an AI model using medical history data from the US as reference. Thereafter, we applied the model to medical history data from Japan and verified its efficacy using a Japanese dataset. Note that we used a large dataset of patients from US as input for the model, considering that HAE is a rare disease.

Initial Model Development with US Dataset

Data selection

The Merative MarketScan Explorys Claims-EMR Dataset (formerly IBM Watson Health) [8] was used to obtain patient-level linked claims and EMR data for US patients. The diagnoses and prescription histories of patients with edema or digestive symptoms from January 2012 to January 2021 were identified from the dataset and were used to build our model. Data from a total of 4,283,815 patients were used in the study.

To identify the diagnosis history of patients, the International Classification of Diseases (ICD) [9] code available in this dataset was used. However, the ICD code for HAE (D84.1) represents “defects in the complement system,” which is also applicable to other similar diseases. Therefore, we used the prescription history of drugs administered only for HAE (Table 1) to distinguish patients with HAE. We categorized the patients with a prescription history of these drugs as the "HAE group," representing patients presumed to have HAE.

To maintain the demographic characteristics of the original data, the control group was randomly sampled from 1% of the remaining patients, with a fixed ratio of age groups and male-to-female ratio (Figure 1). Note that this was crucial to reduce the data volume to operate the model using limited computation resources (2 central processing units and 16 GB of memory). This was done considering the potential utilization of the model in various medical institutions in the future.

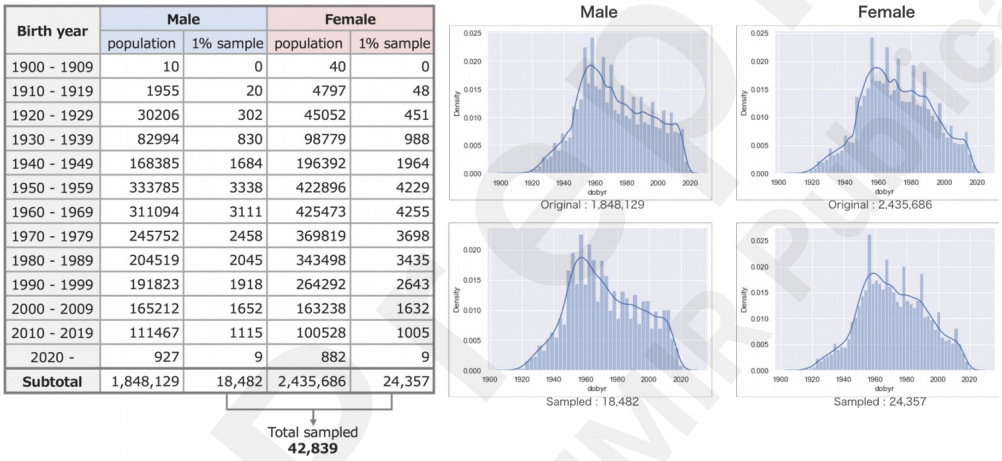
Finally, three groups were included for model development and validation (Figure 2): HAE group with 179 patients; D84.1 (including individuals that likely have HAE but do not having a prescription history of HAE-specific treatments) with 1,521 patients; and control group with 42,839 patients.

To develop the model, the ICD code was used to create features that described the diagnostic history of patients. As this dataset contained both ICD-9 and ICD-10 codes throughout the data period, we standardized the two ICD types. We assigned codes representing the same disease items from both ICD-9 and ICD-10 codes under a single ID.

Table 1. Food and Drug Administration-approved medications used only for HAE (as of January 2022).

Proprietary Name	Nonproprietary Name	Product NDC
BERINERT	HUMAN C1-ESTERASE INHIBITOR	63833-825
CINRYZE	HUMAN C1-ESTERASE INHIBITOR	42227-081 42227-083
FIRAZYR	icatibant acetate	54092-702

HAEGARDA	HUMAN C1-ESTERASE INHIBITOR	63833-828 63833-829
KALBITOR	Ecallantide	47783-101
ORLADEYO	Berotralstat hydrochloride	72769-101 72769-102
RUCONEST	c1 esterase inhibitor recombinant	70383-350 69913-350 71274-350
TAKHZYRO	lanadelumab-flyo	47783-644
ICATIBANT (GENERIC)	icatibant acetate/ ICATIBANT	0093-3066 24201-207 60505-6214 63323-574 68462-828 69097-664 71225-114
SAJAZIR	ICATIBANT	70709-013



Birth year	Male		Female	
	population	1% sample	population	1% sample
1900 - 1909	10	0	40	0
1910 - 1919	1955	20	4797	48
1920 - 1929	30206	302	45052	451
1930 - 1939	82994	830	98779	988
1940 - 1949	168385	1684	196392	1964
1950 - 1959	333785	3338	422896	4229
1960 - 1969	311094	3111	425473	4255
1970 - 1979	245752	2458	369819	3698
1980 - 1989	204519	2045	343498	3435
1990 - 1999	191823	1918	264292	2643
2000 - 2009	165212	1652	163238	1632
2010 - 2019	111467	1115	100528	1005
2020 -	927	9	882	9
Subtotal	1,848,129	18,482	2,435,686	24,357

Figure 1. Comparison of the distribution of the 1% sampled dataset with that of the population.

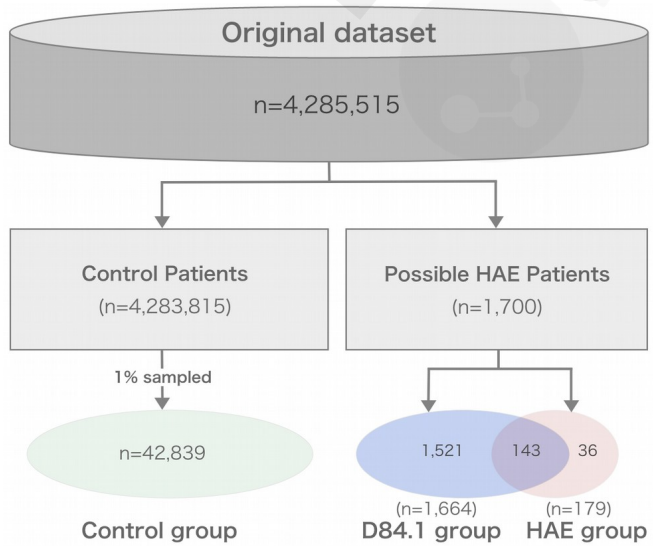


Figure 2. Flowchart depicting the different patient groups created using the US dataset.

Model development

Feature Selection

We counted the number of types of ICD codes diagnosed in both the HAE and D84.1 groups as these two groups should have similar features. Further, the differences in ICD code types between the groups was required to create a model that can identify patients with HAE. We examined rank correlations between the two groups and found it to be approximately 0.08, which suggested that the two groups had different characteristics. We then examined specific ICD codes that were significantly ranked differently between the two groups and identified 25 such ICD codes, which were then used as the primary features in developing the model.

We also examined ICD codes that were diagnosed several times over a period of one year. This is important as patients with HAE tend to have repeated occurrences of swelling in various regions of the body [1], which can lead to the diagnosis of stomachaches and edemas. We counted the number of patients who had been diagnosed with stomachaches or edemas between two to four times per year and found a substantial difference between both groups. Considering that the medical record entry may overlap multiple times when changing the record types, we conducted removal of duplicates based on the date and ICD code for each patient. Thereafter, we labeled a group of ICD codes that related to abdominal pain/edemas and counted the number of times they were assigned in a 1 -year span window for each patient based on this dataset. From this exploratory analysis, we included instances where individuals experienced four or more incidences of stomachaches and three or more incidences of edema per year as part of the main features of our model. The table of the explanatory variables is provided in Multimedia Appendix 1.

Model Building

The number of patients in the HAE group was extremely small compared to that in the control + D84.1 groups; thus, to avoid overfitting, we used bootstrap sampling [12,13] to create the model. A generalized linear model [14] with regularization terms [15,16] was adopted. We used $g(\pi)=\log\left(\frac{\pi}{(1-\pi)}\right)$ for the link function to create a logistic model that would indicate the likelihood of the patient belonging to the HAE group. We chose logistic regression for evaluation as it allows for regression with regularization and is relatively easy to use for evaluating and interpreting feature importance by checking the coefficients. The estimation of the partial regression coefficients was calculated by the maximum likelihood method, which estimates parameters (known as maximum likelihood estimates) that maximize the likelihood of the given observed values. The regularization parameter λ was set to 1 to ensure that it was Lasso regularization.

We used 25% of the data from the HAE group and another 25% from the control + D84.1 groups to train the model, which was then used to predict the remaining 75% of each group. This modeling process was performed 20 times with different random seeds. The average predicted value was calculated as the final output for all the patients. In each trial, the sample used as training data did not have a predicted value and was excluded from the average value calculation (Figure 3). Upon applying the regularization using Lasso regression, the number of substantial features was sorted out

during each calculation by mathematically adjusting coefficients of some variables to 0. The number of sorted features varied with an average of 10; notably, different features were selected every time.

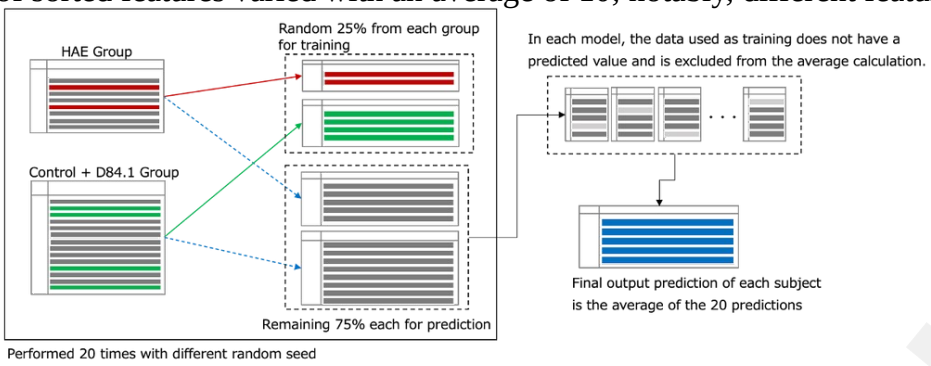


Figure 3. Training data extraction and prediction calculation of the constructed model.

Evaluation Method and Threshold Setting

After obtaining the final value for each participant, we performed Welch's *t*-test on the two distributions to confirm that the two groups had different means. Subsequently, we defined the threshold value that yielded the most balanced classification accuracy using the receiver operating characteristic (ROC) curve. ROC curves are useful for visualizing the entire scenario of trade-offs between sensitivity and precision across a set of cutoff points. The volumes of the HAE and control + D84.1 groups were not equal; therefore, it was important to check the balance between sensitivity and precision rather than the accuracy itself.

Model Application to Japanese EMR Data

Data Extraction and Model Application

For the validation step using Japanese data, data were extracted from a data warehouse (DWH), which collects medical data from the EMR of the KUHP. Patient IDs in the DWH are pseudonymized. The medical data was obtained for a total of 702,213 patients, among which 22 had a history of HAE, 47 had a confirmed diagnosis of HAE, and 123 had a suspected diagnosis of HAE. The data for model validation included those associated with patients from all these groups (patients using drugs for HAE, patients with confirmed HAE, and patients suspected to have HAE). This was done because physicians may have suspected HAE for some patients if their symptoms were similar to those of patients with the condition. Therefore, these three types of patients were considered as the HAE patients in the study (HAE group; Figure 4).

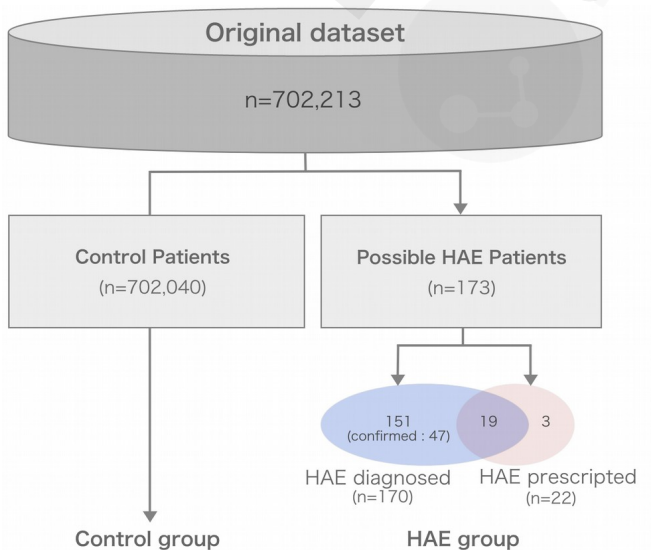


Figure 4. Flowchart depicting the different patient groups obtained from the KUHP dataset.

To adapt the model to Japanese data, we used the standard disease name codes widely used in Japan, as defined by the Medical Information System Development Center (MEDIS-DC) [10],

instead of the ICD code. Although the ICD code is the basic classification code for diagnosis, the standard disease name codes have more subdivisions compared to ICD code, and hence, they can provide a more precise clinical diagnosis. We converted the ICD codes by using the standard disease name code master for ICD-10 [11].

Patient data extracted from the DWH were transferred to Google Cloud Platform server (a virtual private cloud environment) hosted at KUHP. The AI model and statistical programs were stored in a container and sent to the server. We then accessed the server via a virtual private network, which could only be accessed by the authors of the present study. The model was applied to all patient data on this server. This study was approved by the Ethics Committee of the Kyoto University Hospital (approval number: #R3750).

Results

Evaluation of the Initial Model

Welch's *t*-test indicated that the two patient groups did not have same mean values, as suggested by the *p*-value of 2.2e-16. Further, the area under the ROC curve was 86.4%, which was obtained when only the HAE group was set as true and all the other groups as false. The best accuracy threshold of this ROC curve was calculated as 39%, with an accuracy of 99.6%. This is because the volume of the control + D84.1 group was larger than that of the HAE group. The true -positive (sensitivity) of this threshold was only 10.6%, with a precision of 54.3%.

As we aimed to identify patients likely to have HAE, we searched for a different threshold that can improve the sensitivity while keeping the precision at an acceptable level. Considering the fact that the prediction of the HAE group had $\mu \fallingdotseq 0.15$ and $\sigma^2 \fallingdotseq 0.025$, 0.075–0.125 could be a good threshold candidate. We confirmed the sensitivity and precision for the thresholds of 0.075, 0.1, and 0.125 to determine the most balanced threshold, as shown in Figure 5.

The threshold value of 0.1 had a sensitivity of 52.5% and precision of 27.1%, indicating that one out of two known HAE group participants can be correctly detected, and one out of four detected participants should correctly belong to the HAE group. If we exclude the HAE group participants who were not diagnosed with D84.1, the sensitivity was 61.5%. This result was calculated based on 1% of the sample size of the original control group; thus, by multiplying the number of all subjects from the control group by 100, we obtained a 100% scale precision of 2%. This was two times better than the 1% precision goal set at the beginning of the study. This means that based on this model, 1 out of 50 suspected patients is highly likely to have HAE. Considering that HAE prevalence is estimated to be 1 in 50,000 people, we can expect to find undiagnosed HAE patients quickly and efficiently using this model output.

From a conservative standpoint, the threshold value of 0.1 seems to be optimal. However, to identify more potential patients with HAE, it might be better to apply the 0.075 threshold, which has a sensitivity of 55.3% and a precision of 23.9%. If we recalculate the 100% scale precision in the same manner as described above, we obtain 1.6%. This means that we can still achieve our goal of 1% precision while improving the sensitivity.

In addition, we need to consider the fact that the ratio of suspected patients in the US dataset can be calculated to be approximately 0.09% with a 0.1 threshold, and 0.15% with a threshold of 0.075. If

this model is to be used on a much smaller volume dataset compared to the US dataset, there is an approximately two-times higher risk of obtaining zero suspected patients with a 0.1 threshold than with the 0.075 threshold.

Suspection stat (threshold = 0.1)	Control Group		D84.1 Group	HAE Group	
	100% scale converged	1% scale			not D84.1
Not Suspected	4,279,500	42,795	1,312	55	30
Suspected	4,400	44	209	88	6

1% scale precision	27.1%	=94/347	sensitivity*1	52.5%	=94/179
100% scale precision	2.0%	=94/4703	sensitivity*2	61.5%	=88/143

*1: Including "not D84.1" patients

*2: Excluding "not D84.1" patients

Suspection stat (threshold = 0.075)	Control Group		D84.1 Group	HAE Group	
	100% scale converged	1% scale			not D84.1
Not Suspected	4,277,900	42,779	1,266	52	28
Suspected	6,000	60	255	91	8

1% scale precision	23.9%	=99/414	sensitivity*1	55.3%	=99/179
100% scale precision	1.6%	=99/6354	sensitivity*2	63.6%	=91/143

*1: Including "not D84.1" patients

*2: Excluding "not D84.1" patients

Suspection stat (threshold = 0.125)	Control Group		D84.1 Group	HAE Group	
	100% scale converged	1% scale			not D84.1
Not Suspected	4,280,600	42,806	1,343	62	33
Suspected	3,300	33	178	81	3

1% scale precision	28.5%	=84/295	sensitivity*1	46.9%	=84/179
100% scale precision	2.4%	=84/3562	sensitivity*2	56.6%	=81/143

*1: Including "not D84.1" patients

*2: Excluding "not D84.1" patients

Figure 5. Cross-tabulation calculated at three different threshold values using all data groups for detailed evaluation of different scaled precisions and group sensitivities.

Application of the Model to Japanese EMR Data

To verify the performance of this model using Japanese data, it was applied to patient data obtained from KUHP, and the output of potential patients with HAE was obtained based on the selected threshold. The diagnostic histories of these patients were stored at a single university hospital. Compared to the dataset used to build the original model, the variation and coverage of the entire diagnostic history were assumed to be relatively low. Therefore, we adopted a threshold value of 0.075 in this validation study to aggressively identify patients with HAE. We considered the HAE group (Figure 4) as the correct data for this validation.

As shown in Figure 6, 65 of 173 HAE patients were detected using this model, indicating a sensitivity of 37.6%. Some patients in the HAE group did not have a diagnostic history specific to HAE (e.g., abdominal pain, swelling, or edema) within the KUHP data. Their common symptoms might have been treated by their primary doctors/clinics and not at this university hospital. Moreover, because HAE is a hereditary disorder, some patients may have been diagnosed through

family testing. These factors appear to lead to a lower sensitivity score for the Japanese dataset than that for the US data.

The precision score was 23.6%, which is more than 14 times higher than that of the initial model. As mentioned in the Introduction, only 20% of patients in Japan are diagnosed with HAE, which means that 80% of patients with HAE are undiagnosed. Therefore, the 211 patients from the control group who were suspected to have HAE in our model may be undiagnosed HAE patients.

Suspection stat (threshold =0.075)	Control Group	HAE group		
		diagnosed		prescripted
Not Suspected	701,829	93	13	2
Suspected	211	58	6	1

Precision	23.6%	=65/276	sensitivity	37.6%	=65/173

Figure 6. Cross-tabulation with precision and sensitivity scores of KUHP results.

Discussion

Principal Findings

In this study, we developed an AI model for screening patients with HAE and validated its performance using two methods.

First, a large patient dataset was selected to build a model containing patient-level linked claims and EMR data from the US. The advantage of this dataset is that it contains a long-term prescription and diagnostic history across multiple medical institutes. The diagnostic characteristics of patients with HAE were determined by analyzing the dataset. Based on these characteristics, we constructed a generalized linear model with regularization terms. At a threshold of 0.1, the sensitivity score was 52.5% and the precision score was 27.1% if patients with possible HAE were included in the correct answer group. When these were excluded from the correct answers, the sensitivity score was 61.5%.

We then applied this model to Japanese EMR data. This validation was conducted at a single university hospital using DWH data. Generally, patients often visit local hospitals and rarely visit university hospitals if they present with common symptoms. Considering this situation, data obtained from a single university may have some difficulty with model performance. Although the sensitivity score was lower than that of the US dataset (37.6%), the precision score reached 23.6% with a threshold value of0.075. This implies that our model has a high possibility of identifying patients with undiagnosed HAE in Japan.

Limitations

Our study had several limitations. Generally, because HAE is a rare disease, patient group data (correct answer data in machine learning) are quite small. In addition, the variance in each patient's features was larger than that in common diseases. We also suggest possible limitations and countermeasures.

Family History

In our basic analysis of the HAE group, we found that some patients in the HAE group had a lower diagnostic history than the others. We suspected that these patients had been diagnosed with HAE

based on their family histories. Because our model uses the diagnostic history to calculate the probability, these cases are potentially difficult to detect.

US Patient Data Consists of Data from Multiple Hospitals

Our model may rely on the fact that US patient data consist of data from multiple hospitals. Collecting data from multiple hospitals will allow tracking of the records of a single patient across these hospitals and provide a longer/more detailed medical history. For validation in the Japanese dataset, we could only use data from a single university hospital, which may be one of the reasons for low sensitivity.

Potential HAE Patients Might be Included in the Control + D841 Groups

Since the HAE diagnosis rate was low, it is likely that there were more HAE patients in the control + D84.1 groups. In our approach, we assigned the HAE group a prescription history of HAE drugs to keep the model conservative.

Possible Difference in Diagnostic Tendency Between US and Japan

If there are differences in how doctors make diagnostics between countries, we may need to customize the model or threshold to adapt it to Japan and other countries.

Comparison with Prior Work

Few previous studies have focused on screening patients for rare diseases based on diagnostic histories such as medical claims. Nonetheless, some studies have focused on a few rare diseases. For example, a previous study utilized AI models based on diagnostic history to identify patients with Pompe disease [17]. In this study, 104 patients were flagged by the model to have the disease but only 19 were determined by specialists to have a high likelihood of having Pompe disease, rendering a precision score of 18.27% [17]. In comparison, our model recorded a precision of 23.6%. Screening for rare diseases is extremely difficult compared to other common diseases, for which abundant data exist; however, our results indicate that AI models can show high performance for screening rare diseases.

Conclusions and Future Directions

Considering the prevalence of HAE (1/50,000), the screening performance of this model was 1,000 times greater than that achieved through random searching using US data. Owing to their prevalence and recognition rates, identifying undiagnosed patients with rare diseases is an arduous task. This study suggests that patient screening for HAE may become significantly more efficient if this AI model is used. This approach is particularly valuable for the diagnosis and treatment of rare diseases

Additionally, during the validation phase using the Japanese data, the model was effective at a single university hospital. Although only the diagnosis codes recorded in the EMR were available, the model could detect patients with typical symptoms of HAE. The performance of the model can likely be improved further if this model is applied to the data from city hospitals or medical claims, which contain diagnostic histories of patients in multiple medical institutions. This can provide more comprehensive information on the symptoms and diagnostic histories of each patient.

In this study, only patients with a diagnostic history of HAE within the dataset were defined as correct answers. By providing a diagnosis rate, these data may include patients with undetected HAE. The model performance cannot be strictly calculated in such situations. Therefore, further

studies are needed to determine whether patients with undiagnosed HAE should be included in the predicted group. Because identifying undiagnosed patients with HAE is a critical issue, especially in Japan, we will implement a prospective clinical study utilizing our AI model.

The constructed model may help researchers, physicians, and other healthcare professionals identify undiagnosed HAE cases. Eventually, if this strategy can identify undiagnosed patients and provide them with proper treatment, their quality of life will likely be improved.

Acknowledgements

We acknowledge the support received from DISCOVERY (Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema), which covered the publication costs of the study.

Conflicts of Interest

The authors have no financial conflicts of interest.

Abbreviations

HAE: Hereditary Angioedema
EMR: Electronic Medical Record
AI: Artificial Intelligence
KUHP: Kyoto University Hospital
ICD: International Classification of Diseases
NDC: National Drug Code
MEDIS-DC: Medical Information System Development Center
DWH: Data warehouse
ROC: Receiver Operating Characteristic

Multimedia Appendix

Multimedia Appendix 1: [Table of the explanatory variables.]

References

1. Bork K, Meng G, Staubach P, Hardt J. Hereditary angioedema: new findings concerning symptoms, affected organs, and course. Am J Med. 2006 Mar;119(3):267-74. doi: 10.1016/j.amjmed.2005.09.064. PMID: 16490473.

2. Zuraw BL. Clinical practice. Hereditary angioedema. N Engl J Med 2008 Sep 04;359(10):1027-36. doi: 10.1056/NEJMcp0803977. PMID: 18768946.

3. Ohsawa I. Nanbyo Iden-sei kekkann-sei fushu HAE. (An intractable disease: Hereditary Angioedema (HAE)), Osaka: Iyaku Jānarusha, 2016. ISBN: 4753228134 (in Japanese)
4. Iwamoto K, Yamamoto B, Ohsawa I, Honda D, Horiuchi T, Tanaka A, et al. The diagnosis and treatment of hereditary angioedema patients in Japan: A patient reported outcome survey. Allergol Int 2021 Apr;70(2):235-243. doi: 10.1016/j.alit.2020.09.008. Epub 2020 Nov 6. PMID: 33168485.

5. Bellanti JA, Settipane RA. The Floralia: A festive time for Romans and a demanding time for the allergist/immunologist. Allergy Asthma Proc 2018 May 01;39(3):167-168. doi: 10.2500/aap.2018.39.4141. PMID: 29669662; PMCID: PMC5911509.

6. Zanichelli A, Magerl M, Longhurst H, Fabien V, Maurer M. Hereditary angioedema with C1 inhibitor deficiency: delay in diagnosis in Europe. Allergy Asthma Clin Immunol 2013 Aug 12;9(1):29. doi: 10.1186/1710-1492-9-29. PMID: 23937903; PMCID: PMC3751114.

7. DISCOVERY (Diagnostic Consortium to Advance the Ecosystem for Hereditary Angioedema). URL: https://discovery0208.or.jp/en/top/ [accessed 2024-6-7]

8. Merative: Real-world evidence solutions for life sciences. URL: https://www.merative.com/content/dam/merative/documents/brief/real-world-evidence-solution-brief.pdf [accessed 2024-6-7]

9. WHO : The International Classification of Diseases (ICD) URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2024-6-7]

10. Medical Information System Development Center (MEDIS-DC). URL: http://www.medis.or.jp/ [accessed 2024-6-7]

11. Medis: Standard disease name master for ICD-10. URL: http://www2.medis.or.jp/stdcd/byomei/byomei.html [accessed 2024-6-7]

12. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Stat 1979;7(1):1-26.

13. Tibshirani RJ, Bradley E. An introduction to the bootstrap. New York, NY: Chapman & Hall; 1993.

14. McCullagh P, Nelder J. Generalized linear models. 2nd edition. London: Chapman & Hall/CRC; 1989.

15. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser A Stat Soc Series B (Methodological) 1996;58(1):267–288.

16. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970;12(1):55–67.

17. Lin S, Nateqi J, Weingartner-Ortner R, Gruarin S, Marling H, Pilgram V, Lagler FB, Aigner E, Martin AG. An artificial intelligence-based approach for identifying rare disease patients using retrospective electronic health records applied for Pompe disease. Front Neurol. 2023 Apr 21;14:1108222. doi: 10.3389/fneur.2023.1108222. PMID: 37153672; PMCID: PMC10160659.

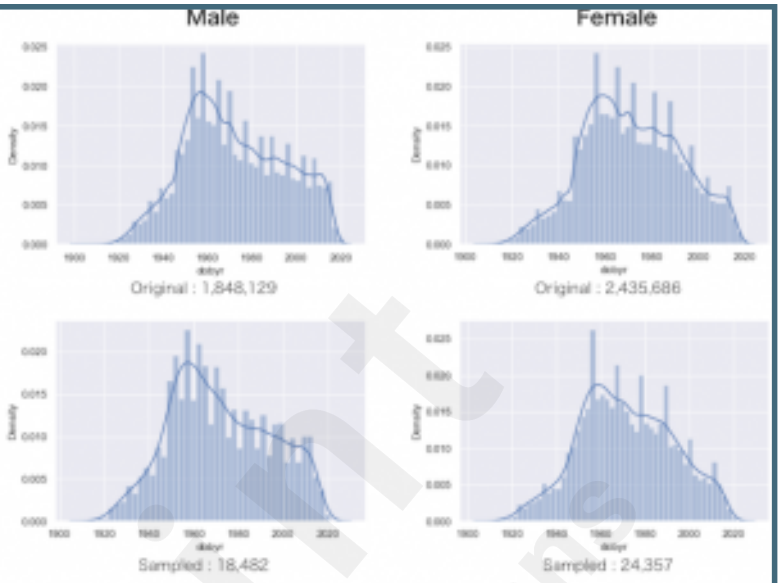
Supplementary Files

Figures

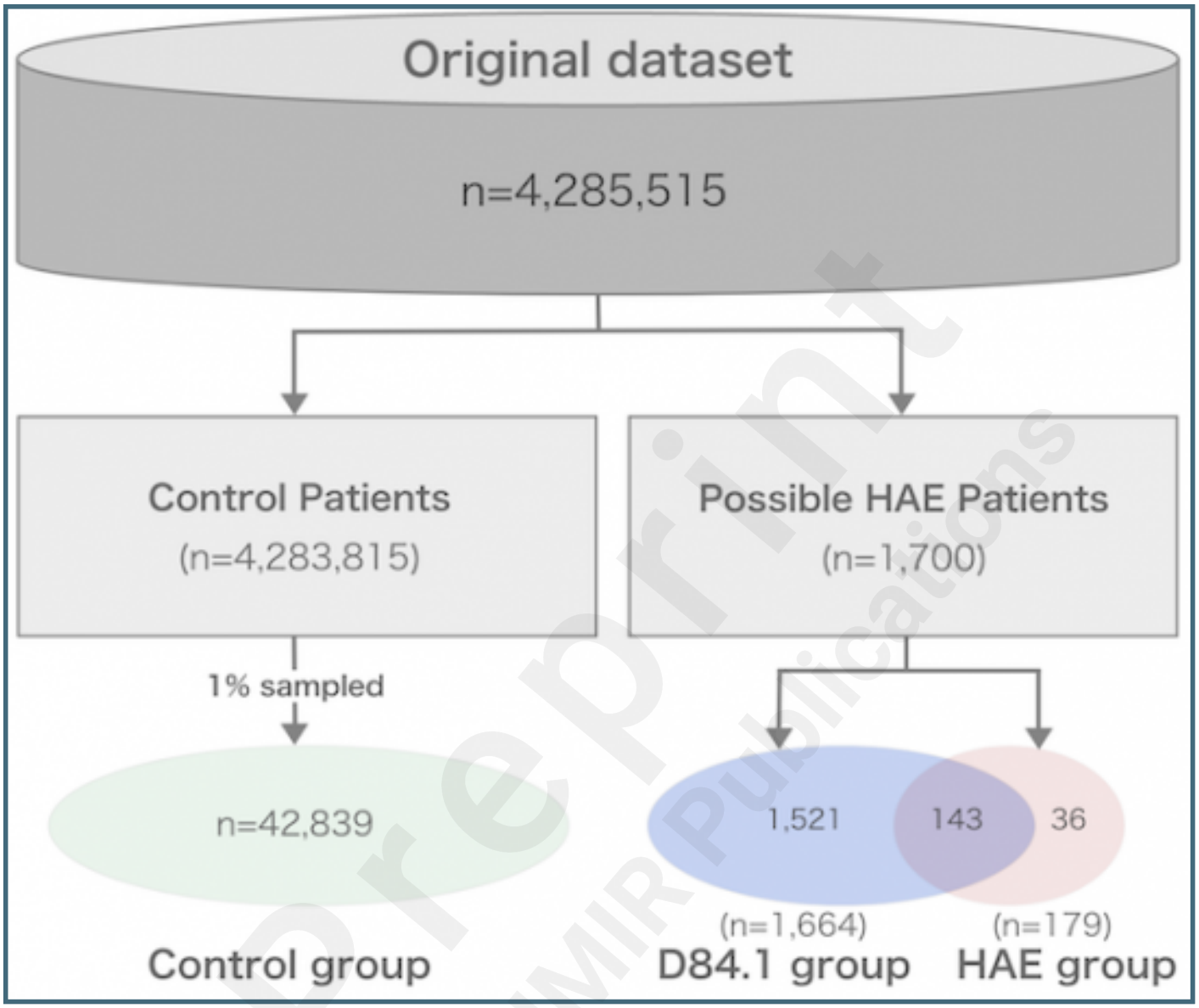
Comparison of the distribution of the 1% sampled dataset with that of the population.

Birth year	Male		Female	
	population	1% sample	population	1% sample
1900 - 1909	10	0	40	0
1910 - 1919	1955	20	4797	48
1920 - 1929	30206	302	45052	451
1930 - 1939	82994	830	98779	988
1940 - 1949	168385	1684	196392	1964
1950 - 1959	333785	3338	422896	4229
1960 - 1969	311094	3111	425473	4255
1970 - 1979	245752	2458	369819	3698
1980 - 1989	204519	2045	343498	3435
1990 - 1999	191823	1918	264292	2643
2000 - 2009	165212	1652	163238	1632
2010 - 2019	111467	1115	100528	1005
2020 -	927	9	882	9
Subtotal	1,848,129	18,482	2,435,686	24,357

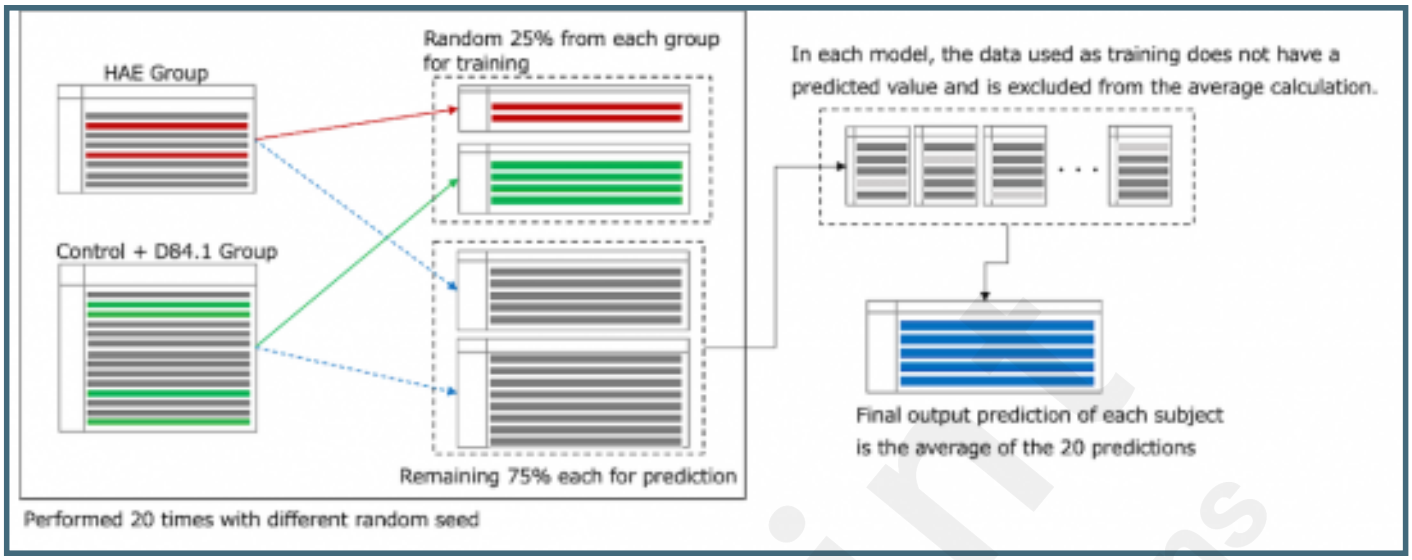
Total sampled
42,839



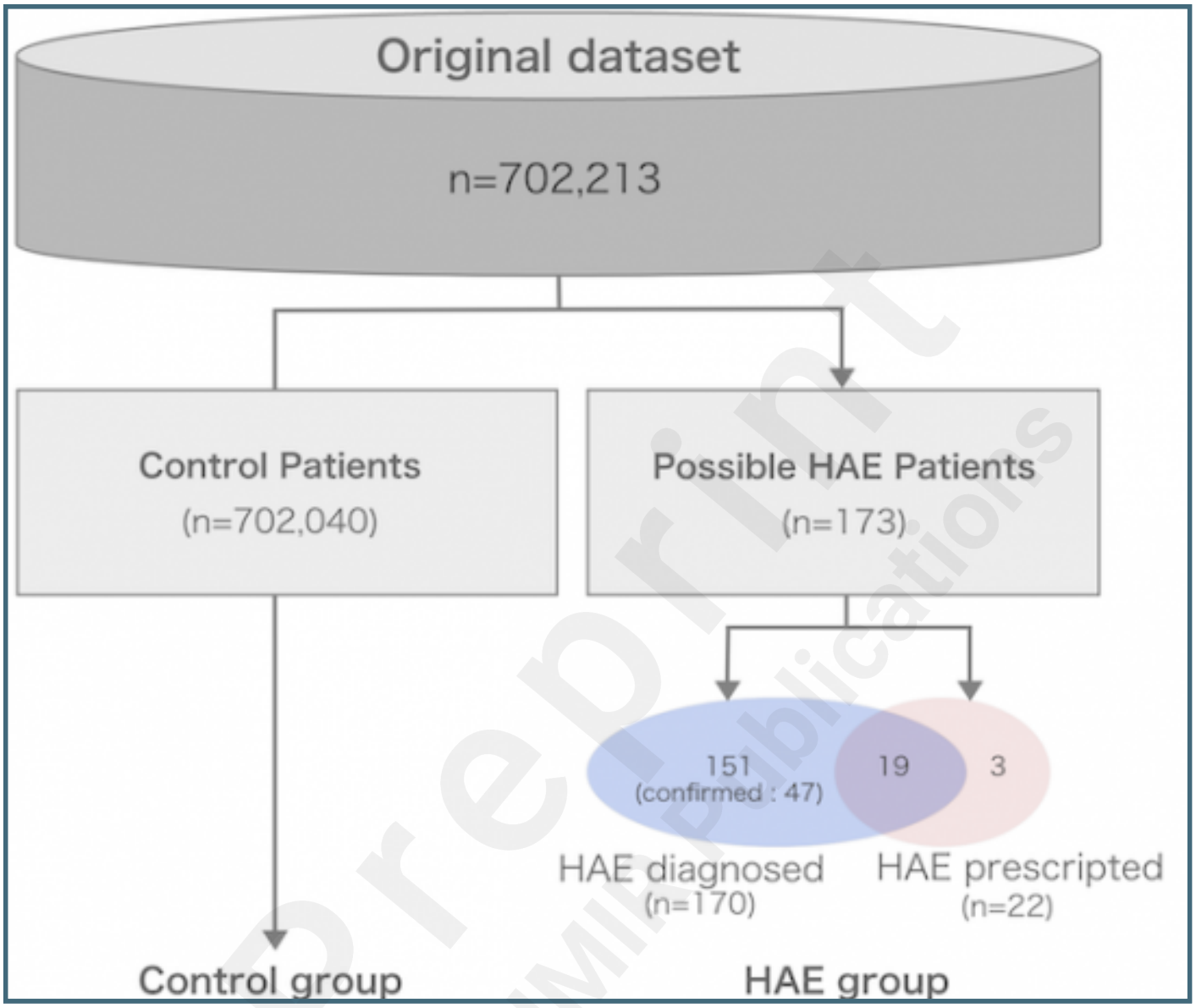
Flowchart depicting the different patient groups created using the US dataset.



Training data extraction and prediction calculation of the constructed model.



Flowchart depicting the different patient groups obtained from the KUHP dataset.



Cross-tabulation calculated at three different threshold values using all data groups for detailed evaluation of different scaled precisions and group sensitivities.

Suspection stat (threshold = 0.1)	Control Group 100% scale converged	Control Group 1% scale	D84.1 Group	HAE Group	HAE Group not D84.1
Not Suspected	4,279,500	42,795	1,312	55	30
Suspected	4,400	44	209	88	6

1% scale precision	27.1%	=94/347	sensitivity*1	52.5%	=94/179
100% scale precision	2.0%	=94/4703	sensitivity*2	61.5%	=88/143

*1: Including "not D84.1" patients

*2: Excluding "not D84.1" patients

Suspection stat (threshold = 0.075)	Control Group 100% scale converged	Control Group 1% scale	D84.1 Group	HAE Group	HAE Group not D84.1
Not Suspected	4,277,900	42,779	1,266	52	28
Suspected	6,000	60	255	91	8

1% scale precision	23.9%	=99/414	sensitivity*1	55.3%	=99/179
100% scale precision	1.6%	=99/6354	sensitivity*2	63.6%	=91/143

*1: Including "not D84.1" patients

*2: Excluding "not D84.1" patients

Suspection stat (threshold = 0.125)	Control Group 100% scale converged	Control Group 1% scale	D84.1 Group	HAE Group	HAE Group not D84.1
Not Suspected	4,280,600	42,806	1,343	62	33
Suspected	3,300	33	178	81	3

1% scale precision	28.5%	=84/295	sensitivity*1	46.9%	=84/179
100% scale precision	2.4%	=84/3562	sensitivity*2	56.6%	=81/143

*1: Including "not D84.1" patients

*2: Excluding "not D84.1" patients

Cross-tabulation with precision and sensitivity scores of KUHP results.

Suspection stat (threshold =0.075)	Control Group	HAE group	
		diagnosed	prescripted
Not Suspected	701,829	93	13	2
Suspected	211	58	6	1

Precision	23.6%	=65/276	sensitivity	37.6%	=65/173

Multimedia Appendixes

Table of the explanatory variables.
URL: http://asset.jmir.pub/assets/2284f7a667b63059313d1248ed6fc64e.png

Related publication(s) - for reviewers eyes onlies

ADDITIONAL MATERIAL (Sensitivity & Precision of all thresholds).