MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)
© 2017 International Medical Informatics Association (IMIA) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-830-3-432

The Impact of "Possible Patients" on Phenotyping Algorithms: Electronic Phenotype Algorithms Can Only Be Reproduced by Sharing Detailed Annotation Criteria

Rina Kagawa, MD[a], Yoshimasa Kawazoe, MD, PhD[b], Emiko Shinohara, PhD[b], Takeshi Imai, PhD[c], Kazuhiko Ohe, MD, PhD[a,b]

[a] *Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Japan*
[b] *Department of Healthcare Information Management, The University of Tokyo Hospital, Japan*
[c] *Center for Disease Biology and Integrative Medicine, The University of Tokyo, Japan*

Abstract

Phenotyping is an automated technique for identifying patients diagnosed with a particular disease based on electronic health records (EHRs). To evaluate phenotyping algorithms, which should be reproducible, the annotation of EHRs as a gold standard is critical. However, we have found that the different types of EHRs cannot be definitively annotated into CASEs or CONTROLs. The influence of such "possible patients" on phenotyping algorithms is unknown. To assess these issues, for four chronic diseases, we annotated EHRs by using information not directly referring to the diseases and developed two types of phenotyping algorithms for each disease. We confirmed that each disease included different types of possible patients. The performance of phenotyping algorithms differed depending on whether possible patients were considered as CASEs, and this was independent of the type of algorithms. Our results indicate that researchers must share annotation criteria for classifying the possible patients to reproduce phenotyping algorithms.

Keywords:

Clinical Phenotyping, Data Annotation, Electronic Health Records

Introduction

Background and problems

To improve healthcare quality and clinical research, it is critical to identify patients diagnosed with a particular disease. As structured data on diagnoses in electronic health records (EHRs) are limited in terms of accuracy and completeness [1, 2], the demand for automated techniques for identifying patients diagnosed with a particular disease based on EHRs, so-called phenotyping, has been increasing [3, 4]. In previous studies, we applied published algorithms to EHR datasets of Japanese patients [5, 6], which required annotated EHRs as a gold standard to evaluate the algorithms. Among the several annotation techniques, we chose manual annotation because it is more accurate than others, and over 75% of previous studies employed it [4]. Through the annotation process, we found that it is often difficult to annotate EHRs as definite CASE or CONTROL patients as a gold standard [6].

Annotation criteria are critical for phenotyping algorithms because they directly affect the calculation of the algorithms' performance. If researchers do not share how to annotate such "possible patients," the published performance of a phenotyping algorithm would differ among research teams because the CASEs would differ, even if each dataset had identical characteristics (Figure 1).

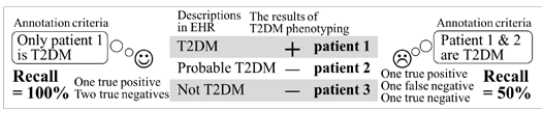

Figure 1– The performance of phenotyping algorithm (recall) for classifying T2DM cannot be reproduced across institutions due to the different annotation criteria for classifying the "possible patient (patient 2)."

Why is the annotation of possible patients indefinite?

In EHR annotation for type 2 diabetes mellitus (T2DM), we identified three types of EHR data in terms of possible patients [6]. Each of these suggests a different likelihood of T2DM, as shown in the following examples:

Example 1: Antiglutamic acid decarboxylase (GAD) antibody ≧ 1.6 U/mL and antiislet antigen 2 (IA2) antibody ≧ 0.4 U/mL

Example 2: *"...T2DM is likely..."*

Example 3: *"...He met the diagnostic criteria for DM... Type 1 DM is unlikely because… Secondary DM is denied..."*

In example 1, the explicit information about the disease name, such as the direct noting of "type 1 DM," is not provided, but the information implies a low possibility of T2DM; that is, the high value of the anti-GAD antibody or anti-IA2 antibody suggests type 1 DM. Example 2 includes an explicit but indefinite description (*"likely"*), and the annotation result may differ among annotators. Example 3 provides no direct information about the type of DM, but the contextual information can increase the conviction that the patient has T2DM; that is, it denies other types of DM. These ambiguous descriptions are sufficient for medical experts who have medical knowledge, and they sometimes dare to describe EHRs ambiguously to accurately record the facts when they cannot diagnose patients with certainty [7]. However, these descriptions are not necessarily sufficient for researchers who sometimes expect the definite truth. This is one of the essential limitations of retrospective EHR-based studies across institutions or countries. Moreover, the fact that each patient's EHRs contain many such ambiguous descriptions makes reproducible annotation difficult.

As surveyed, one study separated definite and possible CASEs to identify rheumatoid arthritis and grouped possible CASEs with CONTROLs [8]. For other diseases, such as multiple sclerosis, the possible patients were identified for multiclass classifications [9–11]. However, they have not shown multiple types of possible patients, and the published annotation criteria were disease dependent. For other diseases, even the existence of possible patients as a gold standard has not been examined. No studies have shown whether the performance of a

phenotyping algorithm will be influenced by which types of possible patients are classified as CASEs. We aim to quantitatively clarify these issues and mitigate the ambiguities to facilitate reproducible phenotyping algorithms [12].

Research objective, novelty, and related work

We analyze EHR data to examine the impacts of the annotation criteria for classifying possible patients on (1) the proportion and characteristics of CASE patients and (2) the performance of phenotyping algorithms. To accomplish this, we create a nonbinary annotation method based on the conviction of a target disease. Our targets are four chronic diseases. This study's originality is to handle directly the uncertainties in records of diseases and their influences on the algorithms' performance using real world data. One study simulated the loss of power of phenotyping algorithms due to bias in the EHR data [13], but did not mention the information bias related to the annotation criteria. No phenotyping studies have addressed the uncertainties in disease records [7].

Methods

Target diseases, eligible subjects, and EHR data

Our targets are four diseases from the Unified Medical Language System (UMLS) Metathesaurus—common diseases: (i) T2DM and (ii) essential hypertension (HT); and rare diseases: (iii) primary biliary cirrhosis (PBC; the full disease name is changed since approximately 2015 [14], but this study used data until 2014) and (iv) autoimmune hemolytic anemia (AIHA). We selected chronic diseases, which have associated clinical guidelines in Japan [15–18], to focus only on the presence of diseases. A disease must be diagnosed through a combination of several tests. If a disease can be diagnosed using one test, it does not require a phenotyping algorithm. This study involved 650 patients (mean age 52.6; 57.2% female) randomly selected out of 104,522 patients who made at least two visits to the University of Tokyo Hospital between 1/1/2009 and 12/31/2014 and at least one visit in 2012. We used the EHR data over six years (2009–2014).

Detailed annotation (DA) method

In this study, annotators checked EHR data retrospectively and determined a "CASE" based on the degree of conviction that the patient had a target disease, which was recorded in EHRs by clinicians who examined the patient. The examples of EHRs in the Introduction section suggest that possible patients should be annotated into several types and that annotation should be independent of the data structure. Labelling the information itself would be useful for reproducible annotation of any EHRs because the annotation results would differ among research teams using different information. We divided the information in the EHRs into two axes, namely explicit information and context (Figure 2); moreover, along each of the two axes, we classified the elements affecting multilevel degree of conviction of a target disease. We have called this the detailed annotation (DA) method (Table 1).

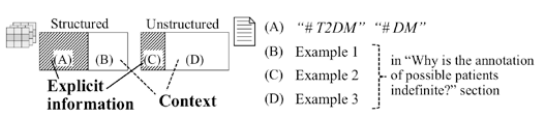

Figure 2– Annotation axes: explicit information and context.

Explicit information includes definite (Table 1(a), (d)) and possible ((b), (e)) descriptions of disease names, and meeting the diagnostic criteria ((c), (e)), which means that annotators can retrospectively determine that the patient met the diagnostic criteria. The description *"T2DM is likely"* is an explicit, possible description of the target disease (T2DM) and belongs to explicit information (b). Contexts include the EHR data implying a target disease (α); upper diseases (β); and the absence ((α), (β)) and presence (ε) of differential diagnoses and sibling diseases; and possible (δ) and definite (ε) descriptions of diseases which are treated by the same medication used for the target disease. For T2DM, the description *"Type 1 DM is unlikely"* belongs to context (α) because T1DM is a sibling disease of T2DM; this description implies T2DM. The pair of the structured data of the high values of the anti-GAD antibody and anti-IA2 antibody belongs to context (ε) because they suggest T1DM. For essential HT, the narrative and definite descriptions that the patient is medicated with an antihypertensive drug not for HT but for angina belongs to (ε); this description implies weak conviction that the patient has essential HT.

The combination of each element of each of the two axes leads to the seven categories of disease conviction, from a definite CASE (category (1)) to a definite CONTROL (category (7)). As an intermediate concept, category (4) indicates the upper disease. A patient with explicit information (b) and context (α) is classified into DA category (2). Patients in the categories (2)–(6) are possible patients.

Experimental setups

Two clinicians (authors) each annotated the EHRs of all 650 patients for each of the four diseases based on the DA categories. The annotators discussed and made final decisions for mismatches. We used Fisher's exact test and Fisher's pairwise exact test (Bonferroni correction) to compare the proportion of possible patients among the diseases. To analyze the patients' characteristics according to the DA categories, we performed statistical analyses of the null hypothesis that the averages of the maximum value of each lab test of each patient or the proportions of each element used in the guideline-based phenotyping algorithms (described later) would be equal among the categories. For continuous data, the variances of each category were found to be unequal by a Bartlett test; the nonparametric Kruskal–Wallis test and the Mann–Whiteney U test (Bonferroni correction) for pairwise comparisons were performed. For categorical data, we used Fisher's exact tests and Fisher's pairwise exact test (Bonferroni correction). The threshold for significance was $p<0.05$. We used R-3.1.3, the coin package 1.1-2, and the fmsb package 0.5.2.

To assess the impact of the annotation criteria on the performance of phenotyping algorithms, authers developed rule-based phenotyping algorithms for each disease (Table 2) using data that are referred in each guideline [15–18] and are stored according to HL7 2.5 (ISO 27931:2009). We call these guideline-based (GB) algorithms. For each algorithm, we changed the annotation criteria by altering the threshold of the DA category in which patients are considered as CASEs from the category (1) to (6), and calculated the values of the evaluation metrics at each threshold. The DA category (7) was excluded from the threshold because it was an unrealistic assumption for all patients (the categories (1)–(7)) to be CASEs. The evaluation metrics are as follows: recall is True Positive (TP)/(TP+False Negative (FN)), precision was TP/(TP+False Positive (FP)), specificity was True Negative (TN)/(TN+FP), and negative predictive value (NPV) was TN/(TN+FN). To examine whether the impact of the annotation criteria depended on a particular phenotyping algorithm, we performed the same experiment for the other phenotyping algorithms, which used only one billing code for each disease. We call these naïve algorithms. More detailed explanations of the DA method and phenotyping algorithms are presented on GitHub.[1]

Table 1– The DA method. Degrees of disease conviction are classified into categories (1)–(7) indicating certainty, probability, possibility, upper disease, the possibility of other diseases, the probability of other diseases, and definite other diseases respectively.

Explicit information			Context (α) Increasing the conviction of the target disease	(β) Probable upper disease AND no sibling diseases or differential diagnoses	(γ) No other context	(δ) Possible diseases which are treated by the same medications used for the target disease	(ε) Decreasing the conviction of the target disease	(ζ) Definite denial of the target disease
Diagnostic criteria		Description						
Target disease name		(a) Definite	(1)	(1)	(1)	(1)	(3)	Contradiction
		(b) Possible	(2)	(3)	(4)	(5)	(5)	(7)
	(c) Meet		(3)	(4)	(4)	(5)	(6)	(7)
Upper disease name		(d) Definite	(2)	(4) or (6) [2]	(4) or (6) [2]	(5)	(6) or (7) [2]	(7)
	(e) Meet	(e) Possible	(3)	(4) or (6) [2]	(4) or (6) [2]	(5) or (6) [2]	(6) or (7) [2]	(7)
Both target and upper	(f) None of the others		(5)	(6)	(7)	(7)	(7)	(7)
		(g) Denial	Contradiction	(7)	(7)	(7)	(7)	(7)

Table 2– GB phenotyping algorithms developed in this study.

(i) T2DM: Patients with (A) AND ((B) OR (C)), modified [19]	
(A)	Excluding other types of DM
(B)	Antidiabetic medication
(C)	T2DM billing codes AND abnormal lab test more than two times.
(ii) Essential HT: Patients with (A) AND (B)	
(A)	Excluding secondary HT
(B)	HT billing codes except (A) OR medication of ARB/ACE inhibitor
(iii) PBC: Patients with (A) AND (B)	
(A)	Antimitochondrial antibody (AMA) is positive
(B)	(γGTP≧ 68 IU/L AND ALP≧ 359 IU/L) OR PBC billing codes
(iv) AIHA: Patients with ((A) AND (B)) OR (C))	
(A)	More than four abnormal lab tests, which mean hemolytic anemia
(B)	Direct Coombs test is positive AND (AIHA billing codes OR no billing codes for other diseases that cause anemia)
(C)	Disease names of AIHA in EHR, applying the technique in [20]

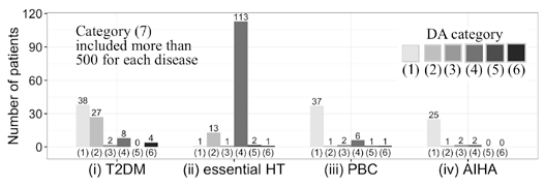

Figure 3– The numbers and distributions of patients classified into each DA category differ among the diseases.

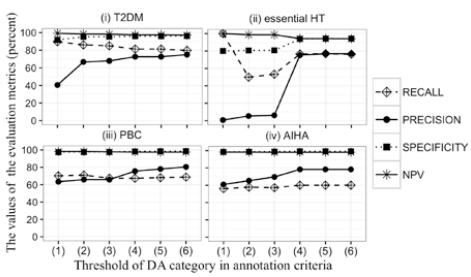

Figure 4– Changes in the performance of GB algorithms depending on the threshold of CASEs differ among the diseases.

Results

Figure 3 shows the annotation results based on the DA categories (weighted κ statistics = 1.00 for each disease). The mismatches between the annotations of two patients was caused by an oversight of the EHRs and those of seven patients were caused by misunderstanding of the contexts. The category (7) included more than 500 patients for each of the four diseases and is excluded from Figure 3. All four diseases included possible patients classified into DA categories (2)–(6). The distributions of patients differed among the diseases. T2DM had a broad peak at the categories (1) and (2), and essential HT had a peak at upper disease (the category (4)). PBC included patients in all DA categories, while AIHA did not include patients in the categories (5)–(6). The proportions of possible patients to all 650 patients were not equal among the diseases (p<2.2e-16, two sided). Post-hoc test showed significant differences between all disease pairs (p<0.001) except the pair of PBC and AIHA (p=0.14).

If the characteristics of patients in the DA categories (1)–(6) were exactly equal, the degrees of disease conviction recorded by the clinicians in EHRs would be completely random, and the classification of possible patients would not be necessary. We performed statistical analyses to assess this. Because the category (7) indicates definite CONTROLs, it will naturally have different characteristics from the other categories and was excluded from the statistical analysis. The averages of the maximum HbA_{1c} of each patient differed significantly among the DA categories (1)–(6) (Table 3(a)). Post-hoc tests showed no significant difference for each category pair (Table 3(b)). The other lab tests' values did not differ significantly. The proportions of tests for glucose (p=0.048), tests for anti-GAD antibody (p=0.021), antidiabetic medication (p=0.0009), insulin (p=0.040), T2DM billing codes (p=0.009), and T1DM billing codes (p=0.0002) differed significantly (two sided) among the categories (1)–(6). A post-hoc test showed no significant differences. Although not significant, the other categorical variables tended to exhibit higher proportions in the category (1) than that in the categories (2)–(6) collectively, i.e., PBC billing code and no billing codes for malignant neoplasms that cause anemia (regarding AIHA, Table 3(c)). [1,2]

Figure 4 shows that the performance of each GB phenotyping algorithm changed depending on the threshold of the DA category in which patients were considered as CASEs. The values of the evaluation metrics at the threshold DA category (2) (hereinafter called th(2)) indicated the values when the definite CASEs (category (1)) and probable CASEs (category (2)) were considered as CASEs. For essential HT, the recall decreased by 50% as the threshold moved from th(1) to th(2) (Figure 4(ii)). It is because that the GB algorithm(ii)(A) excluded 5 out of 13 patients in DA category (2) who had secondary HT billing codes, and that the GB algorithm(ii)(B) also excluded two patients in the category (2) (Table 2). The precision and the specificity increased by 69.2% and 13.1% from th(3) to th(4), respectively; these changes were not found in other diseases. They depended on the 90-patient increase in TP patients and concominant decrease in FP patients, which were due to the peak at category (4) (Figure 3(ii)). In contrast, as the threshold moved from th(1) to th(6), the decrease in the NPV of PBC (0.66%) or AIHA (0.16%), and the increase in the specificity of AIHA (0.63%) were within 1% (Figure 4(iii), (iv)). PBC and

[1] https://github.com/rinabouk/medinfo2017

[2] T2DM and essential HT are smaller in number; they account for more than 50% of upper diseases. PBC and AIHA are larger in number; they account for less than 50%.

AIHA included significantly fewer patients in categories (2)–(6). From th(1) to th(6), the changes in the numbers of TP, FP, FN, and TN patients and the changes in performance were smaller than for common diseases. The changes for T2DM were intermediate; the recall and NPV decreased by 9.73% and 2.12%, respectively, and the precision and the specificity increased by 39.3% and 4.49%, respectively (Figure 4 (i)). Figure 5 shows that the changes in performance of naïve algorithms, which exhibited the same patterns as in Figure 4 except for the decrease in recall from th(1) to th(2) for essential HT.

Table 3(a)– The averages of the maximum HbA_{1c} value of each patient were not equal in the DA categories (1)–(6).

	(1)	(2)	(3)	(4)	(6)	*p*-value
Maximum value of HbA_{1c} of each patient average (SD)	7.78 (1.21) (*n*=35)	7.15 (0.85) (*n*=22)	8.20 (1.40) (*n*=2)	6.76 (0.16) (*n*=5)	6.20 (0.22) (*n*=3)	**0.007**

Table 3(b)– No significant differences between any two category pairs for the maximum HbA_{1c}.

p-value	Effect size	*p*-value	Effect size
(1) (2) 0.28	0.20 (small)	(2) (4) 1.00	0.29 (small)
(1) (3) 1.00	0.02 (no)	(2) (6) 0.17	**0.56 (large)**
(1) (4) 0.33	**0.42 (medium)**	(3) (4) 1.00	**0.82 (large)**
(1) (6) 0.08	**0.57 (large)**	(3) (6) 0.75	**0.97 (large)**
(2) (3) 1.00	0.14 (small)	(4) (6) 0.25	**0.70 (large)**

Table 3(c)– Proportions of billing codes for malignant neoplasms that cause anemia in category (1) (11/14) tended to be lower than in categories (2)–(6) (3/1).

	(1)	(2)	(3)	(4)	(5)	(6)	*p*-value
Patients with billing codes	11	1	0	2	0	0	0.21
Patients with no billing codes	14	0	1	0	0	0	

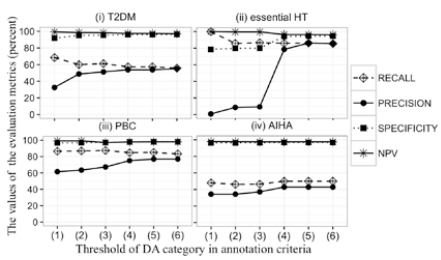

Figure 5– Pattern of changes in the performance of naïve phenotyping algorithms. These are the same as in Figure 4.

Discussion

Our experiments showed that for each of the four chronic diseases, it was necessary to determine how to annotate CASEs by dividing possible patients from definite CASEs and classifying each type of possible patients, to avoid research teams' inclusion of different characteristics in their CASEs. Consequently, the changes in performance of phenotyping algorithms following the alteration in annotation criteria differed among the diseases; this was considered to be independent of the type of algorithms, that is, GB or naïve algorithms. We suggest that these results support the importance of sharing annotation criteria in detail to reproduce algorithms.

The characteristics of possible patients

The different distributions of patients among the diseases (Figure 3) are clinically plausible. Clinicians can diagnose T2DM willingly by a combination of simple lab tests or symptoms [15]; many clinicians were assumed to diagnose T2DM with strong conviction (category (1)) or describe information that inferred T2DM (category (2)). Essential HT is a diagnosis by the exclusion of secondary HT [16]; most clinicians only describe "HT" when referring to essential HT. This was assumed to be the reason for the peak at category (4). PBC and AIHA are rare diseases; then, specialists diagnose most patients with strong conviction (category (1)) [17, 18]. Further, the upper disease of AIHA (acquired hemolytic anemia) is rare, while that of PBC (fibrosis and cirrhosis of liver) cannot be sometimes diagnosed with certainty even by specialists because it is relatively common; then, PBC had a relatively higher proportion of possible patients compared to AIHA, and AIHA did not include the patients in categories (5) and (6).

The detailed annotation criteria according to the degrees of disease conviction recorded by the clinicians reflect patient characteristics and clinicians' rational assessments (Tables 3(a)–(c)). We considered that clinicians could neither diagnose patients with low HbA_{1c} certainly nor describe the definite disease name for such patients, and the average of the maximum HbA_{1c} of each patient in DA categories (4) or (6) for T2DM was relatively low. Similarly, it is suggested that the clinicians' choices of lab tests, medications, or billing codes are affected by the clinicians' conviction of the diseases. It seems one reason no elements tended to be unequal among the categories for essential HT was that they included antihypertensive medications for other diseases and the corresponding billing codes, independent of the conviction of essential HT.

Annotation criteria influence the reproducibility of phenotyping algorithms

Many studies have applied phenotyping algorithms to multiple institutions and identical performance has not been achieved [19, 21]. Our findings suggest that even if the research teams used the same annotation criteria, the ambiguities of the criteria and the corresponding different annotation results could lead to the different performance. This is a critical limitation when other teams attempt to reproduce published performance for different datasets. Without detailed annotation criteria, other teams could not judge whether the different performance arises from differences in annotation criteria, the different characteristics of the study population, or the differences in available data or tools. Our results can guide researchers on this limitation; the robust metrics in terms of changing annotation criteria, such as the NPV of rare diseases, would be preferable for reproducing algorithms without shared annotation criteria. In addition, failure to share annotation criteria could lead to erroneous interpretations of published performance. For example, the precisions of the GB algorithm for T2DM were 39.3% different between th(1) and th(6) (Figure 4(i)).

We confirmed that these issues do not depend on the particular phenotyping algorithm; although only for essential HT, the decrease in recall by 50% from th(1) to th(2) was dependent on the GB algorithm. One simulation study showed that different gold standards led to the different sensitivity or specificity in diagnostic studies [22]. Our results were consistent with this, and presented the first assessments of the different gold standards in studies for phenotyping algorithms using actual data. If all patients diagnosed with a certain disease had definite descriptions of the disease in their EHRs, no patients would be in the categories (2)–(6). Each evaluation metric would retain the same value independent of our annotation criteria, and there would be four flat lines in Figure 4 or 5. The diseases examined in this study were not such cases. Higher recalls of essential HT and PBC, and precision of essential HT could be achieved using naïve algorithms, while the scope of this study was the change in performance. Which algorithm is better is to be determined based on each research objective or available data.

Because the annotation criteria are to be determined according to the research purposes, they will naturally differ among studies. Thus, sharing annotation criteria in detail is critical in reproducing EHR-based studies.

Limitations and future work

To assess our results' generalizability, we must evaluate them with patients with other diseases at other hospitals in several countries. Lower weighted κ statistics would be obtained from this study. Nevertheless, when applying the same phenotyping algorithm to different datasets, the strength of our DA method will not change because it clearly shows that the different results will arise from the different characteristics of the study population, or the differences in available data or tools. Assessment of the impact of clinicians' diagnostic errors is outside the scope of this study and must be done in the future work. We aim to evaluate phenotyping algorithms developed using other techniques. For machine learning, one study indicated that the precision was almost unchanged but the recall differed according to the different training data [21]; thus, different results would be obtained from this study. We will report on which elements used in the DA method can be extracted automatically in the near future. This will lead to a systematic strategy for the development of phenotyping algorithms [12].

Conclusion

Our results confirmed that if researchers do not share annotation criteria in detail for classifying possible patients separately from definite CASEs and for classifying each type of possible patients, the characteristics of CASEs would differ among research teams; although phenotyping algorithms emphasize reproducibility, we cannot expect reproducible performance of the phenotyping algorithm. This was clinically rational for the four chronic diseases. In this study, we annotated EHRs using the DA method. This could increase reproducibility of retrospective EHR-based studies because it achieves annotations with low ambiguity by using information not directly referring to target diseases. We expect that our results will guide researchers on the reproducibility of EHR-based studies.

Ethics and acknowledgements

This research was approved by the Research Ethics Committee of the Graduate School of Medicine and Faculty of Medicine, The University of Tokyo (Permission number: 10733 (2015)). This work was supported by JSPS KAKENHI Grant Number 16J05555.

References

[1] R. Woodfield *et al.,* Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: a systematic review from UK biobank stroke outcomes group, *PLOS ONE* **10** (2015), e0140533
[2] N. McCormick *et al.,* Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis, *PLOS ONE* **9** (2014), e104519
[3] J.C. Kirby *et al.,* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, *JAMIA* **23** (2016), 1046-1052
[4] C. Shivade *et al.,* A review of approaches to identifying patient phenotype cohorts using electronic health records, *JAMIA* **21** (2014), 221-230
[5] R. Kagawa *et al.,* Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach, *J Diabetes Sci Technol* (2016), doi: 10.1177/1932296816681584
[6] R. Kagawa *et al.,* The technical problems for automated phenotyping, *Proc 30th JSAI* (2016), 1-4 (in Japanese)
[7] D.M. Eddy, Variations in physician practice: the role of uncertainty, *Health Affairs* **2** (1984), 74-89
[8] R.J.Carrol *et al.,* Naive electronic health record phenotype identification for rheumatoid arthritis, *AMIA Annu Symp Proc 2011* (2011), 505-511
[9] M.D. Ritchie *et al.,* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record, *JAHG* (2010), 560-572
[10] T. Lingren, Autism data dictionary. Available from: https://phekb.org/sites/phenotype/autism [cited 2016 Dec 9]
[11] S. Bielinski, Heart failure with differentiation between reduced and preserved ejection fraction phenotype algorithm pseudo code cohort version. Available from: https://phekb.org/sites/phenotype/files/HF_algorithm_Cohort.pdf [cited 2016 Dec 9]
[12] R. Richesson and M. Smerek, Electronic health records-based phenotyping. Available from: https://sites.duke.edu/rethinkingclinicaltrials/ehr-phenotyping/ [cited 2016 Dec 9]
[13] R. Duan *et al.,* An empirical study for impacts of measurement errors on EHR based association studies. *AMIA Annu Symp Proc 2016* (2016), 1764-1773
[14] U. Beuers *et al.,* Changing nomenclature for PBC: from 'cirrhosis' to 'cholangitis', *Hepatology* **62** (2015), 1620-1622
[15] The Japan Diabetes Society, Evidence-based practice guideline for treatment of diabetes in Japan, Available from: http://www.jds.or.jp/modules/en/index.php?content_id=44 [cited 2016 Dec 9]
[16] The Japanese Society of Hypertension, Guidelines for the management of hypertension 2014, 2014 (in Japanese)
[17] Research Group on Intractable Liver or Biliary Tract Disease, Clinical guideline for primary biliary cirrhosis. Available from: http://www.nanbyou.or.jp/upload_files/PBC_guideline.pdf [cited 2016 Dec 9] (in Japanese)
[18] Research Group on Idiopathic Hematopoietic Disorder, Reference guideline for clinical practice of autoimmune hemolytic anemia. Available from: http://zoketsushogaihan.com/file/guideline_H26/AIHA.pdf [cited 2016 Dec 9] (in Japanese)
[19] A.N. Kho *et al.,* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study, *JAMIA* **19** (2012), 212-218
[20] A. Turchin *et al.,* DITTO – a tool for identification of patient cohorts from the text of clinician notes in the electronic medical record, *AMIA Annu Symp Proc 2005* (2005), 744-748
[21] R.J. Carrol *et al.,* Portability of an algorithm to identify rheumatoid arthritis in electronic health records, *JAMIA* **19** (2012), e162-e169
[22] A. Karch *et al.,* Partial verification bias and incorporation bias affected accuracy estimates of diagnostic studies for biomarkers that were part of an existing composite gold standard, *J Clin Epidemiol* **78** (2016), 73-82

Address for correspondence
Rina Kagawa, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8655 Japan
+81-3-5800-6427 kagawa-r@m.u-tokyo.ac.jp