International Journal of Medical Informatics 124 (2019) 90–96

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Bias of Inaccurate Disease Mentions in Electronic Health Record-based Phenotyping

Rina Kagawa[1,2,*], Emiko Shinohara[3], Takeshi Imai[4], Yoshimasa Kawazoe[3], Kazuhiko Ohe[2]

[1] Department of Medical Informatics, Strategic Planning, and Management, University of Tsukuba Hospital, Japan
[2] Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, Japan
[3] Department of Artificial Intelligence in Healthcare, Graduate School of Medicine, The University of Tokyo, Japan
[4] Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Japan

A B S T R A C T

Objectives: Electronic health record (EHR)-based phenotyping is an automated technique for identifying patients diagnosed with a particular disease using EHR data. However, EHR-based phenotyping has difficulties in achieving satisfactorily high performance because clinical notes include disease mentions that ultimately signify something other than the patient's diagnosis (such as differential diagnosis or screening). Our objective is to quantify the influence of such disease mentions on EHR-based phenotyping performance.

Methods: Physicians manually reviewed whether the disease mentions indicated the patients' diseases in 487,300 clinical notes of 4,430 patients. Particular focus was placed on disease mentions that did not signify the patient's diagnosis even though they did not have any syntactic modifier or indicator in the same sentences. Patients were then classified according to whether their clinical notes included such disease mentions.

Results: Among the patients whose clinical notes included disease mentions without any modifier or indicator, the proportion of patients whose disease mentions signified the patients' diagnosis was 78.1% (on average). This value can be interpreted as the bias of disease mentions that did not signify the patient's diagnosis on the precision of EHR-based phenotyping by extracting disease mentions from clinical notes.

Conclusion: This study quantified the bias occurred owing to disease mentions that incorrectly signify a patient's diagnosis in the value of precision of EHR-based phenotyping from four dataset types. The results of this study will help researchers in diverse research environments with different available data types.

1. Introduction

The adoption of electronic health records (EHRs) has resulted in easy access and aggregation of clinical data, and increased researchers' interest in using data collected during clinical care [1–3]. One of the most useful pieces of information for the secondary use of EHRs is diagnosis. Since the billing code does not correctly represent a patient's diagnosis [4–6], multiple types of EHR data are necessary to correctly identify a diagnosis; this is known as EHR-based phenotyping and has been widely studied [7–15]. Among the available techniques (such as machine learning or natural language processing (NLP) [16,17]), the most common approach adopted for EHR-based phenotyping is extracting disease mentions (Table 1(A)) from clinical notes [11,18]. This is because clinical notes are considered the most valuable source of clinical information [19–21]. However, achieving high performance (e.g., 95% precision) is difficult [7] for the following reasons: (1) Some clinical notes that do not include disease mentions describe the disease [22], and (2) the presence of some disease mentions does not signify that the patient has a particular disorder [11]. Our previous study focused on the first reason [23]. The present study focused on the second reason. Unlike studies that focused only on problem lists that did not signify the patients' diagnoses [24,25], this study focused on all disease mentions described throughout the free text in clinical notes.

According to previous studies [22,26], disease mentions that did not connote the patients' diagnosis were those with modifiers or indicators in the same sentence (Table 1(C)). However, there are some other disease mentions that do not connote the patient's diagnosis, even though they do not have any modifier or indicator in the same sentences. For example, *prostate cancer* in Table 1(B) was written as a differential diagnosis; the patient did not actually suffer from prostate cancer. Such disease mentions could influence the performance of EHR-based phenotyping by extracting disease mentions from clinical notes. However, the presence of such disease mentions and their influence on EHR-based phenotyping has not been quantified.

The objectives of this study are as follows: (1) To clarify the presence of disease mentions that do not connote the patient's diagnosis, even though they do not have any modifier or indicator in the same sentences. (2) To quantify the influence of such disease mentions on

* Corresponding author at: Department of Medical Informatics, Strategic Planning, and Management, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-8575, Japan.
E-mail address: kagawa-r@md.tsukuba.ac.jp (R. Kagawa).

https://doi.org/10.1016/j.ijmedinf.2018.12.004
Received 7 September 2018; Received in revised form 13 November 2018; Accepted 12 December 2018

Table 1
Examples of the two polarities of disease mentions. These were originally written in Japanese but are provided in English for explanation. The considered disease mentions are underlined. The two principles of polarity are defined in 2.2 Methods. (A) The disease mention signifies the patient's diagnosis. (B, C) The disease mentions do not signify the patient's diagnosis.

	Examples of the disease mentions (Explanations of the polarity of the disease mention)	Principle of polarity	
		syntactic	semantic
(A)	*The patient has right breast cancer.* (There are no indicators or modifiers, and "breast cancer" indicates the patient's diagnosis.)	**positive**	**positive**
(B)	*Prostate cancer* *Atypical adenomatous hyperplasia.* *No tumor found.* (There are no indicators or modifiers, but "prostate cancer" implies one of the differential diagnoses of the patient before the biopsy. "Prostate cancer" does not indicate the patient's diagnosis.)	**positive**	**negative**
(C)	*Mother of the patient has type 2 diabetes mellitus.* ("Mother" is the subject modifier, which implies that type 2 diabetes mellitus is not the patient's diagnosis.)	**negative**	**negative**

EHR-based phenotyping performance. In addition, we classified physicians' intentions to write such disease mentions

2. Materials and methods

2.1. Object diseases and patients

In order to analyze as many disease types as possible from the viewpoints of onset type, diagnostic criteria, and morbidity, we selected the following diseases as the research objects: type 2 diabetes mellitus (T2DM), essential hypertension (HT), diabetic nephropathy (DN), primary biliary cirrhosis (PBC), Crohn's disease (CROHN), prostate cancer (PROSTATE), breast cancer (BREAST), renal cell carcinoma (RCC), dissecting aortic aneurysm (DAA), and pulmonary embolism (PE). This study analyzed 4,430 patients (mean age: 52.6; gender: 57.2% female), randomly selected from patients who had 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 (Appendix A)(Appendix B).

Clinical notes comprised progress notes; summaries; consultation notes; and radiology, endoscopy, and pathology reports written by physicians. The total number of clinical notes was 487,300. To examine the differences depending on the type of clinical notes, we distinguished the discharge summaries and other clinical notes, which were then further classified as semi-structured problem lists or narrative notes (Fig. 1, Table 2).

2.2. Methods

The following experiments were conducted for 10 object diseases and four dataset types.

To perform quantitative manual review of disease mentions that did not signify the patient's diagnosis even though they did not have any modifier or indicator in the same sentences, two principles of polarity of disease mentions were introduced: syntactic and semantic (Table 1). In natural language processing, *polarity* usually refers to character with two poles. In this paper, we use polarity as a framework for discussing features that either do or do not exist, especially regarding diagnoses. According to syntactic polarity, disease mentions without any modifier or indicator (indicating that the disease mention does not signify the patient's diagnosis) are annotated as positive [22,26]. According to semantic polarity, disease mentions are annotated as positive when physicians determine that the disease mention signifies the patient's diagnosis.

2.2.1. Process of quantifying the disease mentions and polarity

Two physicians participated in the following three review steps for each disease (Appendix C). Fig. 2 provides examples of when the dataset type was (Summary-st) in Table 2.

2.2.1.1. Step 1: Identifying disease mentions of the object disease. Similar to the annotation guidelines for disease mentions in clinical notes [22,26], we defined a disease mention as a noun or noun phrase implying the instance of the object diseases. Disease mentions included abbreviations and descriptions in languages other than Japanese, and simple misspellings (e.g., both *type 2 diabetes mellitus*, *DM (type 2)*, and *type 2 diabete mellitus* were considered disease mentions of T2DM).

2.2.1.2. Step 2-a: Annotation of the syntactic polarity of disease mentions. Two existing annotation guidelines mention the modifiers or indicators related to disease mentions. We adopted Elhadad et al.'s subject or generic modifiers (which show that the disease mention does "not pertain to the patient [22]"), negation or uncertainty indicators [22], and Aramaki et al.'s negation, suspicion, and family indicators [26].

The physicians annotated disease mentions with subject or generic modifiers in the same sentence as negative. In this study, a sentence was defined as a text string split by a period or a line break because the clinical notes written in Japanese contained many bullets.

2.2.1.3. Step 2-b: Annotation of the semantic polarity of disease mentions. Disease mentions were annotated as positive when the physician determined that the disease mention indicated the patient's diagnosis. Note that, the diagnosis consisted of patients' current and past medical history. Physicians manually reviewed whether each disease mention indicated the patient's diagnosis based on the physician's inference. For example, if the physician judged that the disease mention did not refer to the patient's diagnosis but rather to a differential diagnosis, it was annotated as negative even if no modifiers or indicators were in the same sentence.

2.2.1.4. Step 3: Quantification of the number of patients classified as positive. We compared the number of patients whose clinical notes included disease mentions annotated as positive using syntactic polarity (Step 2-a) and semantic polarity (Step 2-b); we titled the corresponding patients as (a) patients classified as syntactic positive and (b) patients

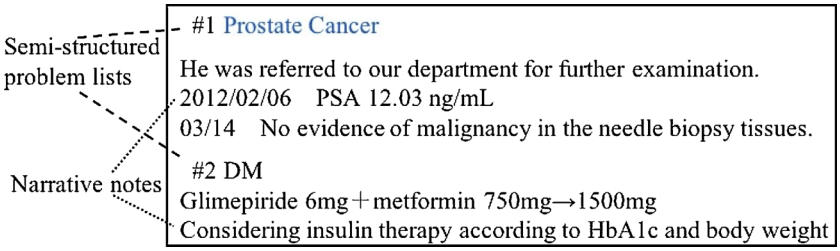

Fig. 1. Examples of clinical notes in EHR (originally written in Japanese but provided in English here for explanation). The blue text *prostate cancer* implies the suspicion of prostate cancer before its final diagnosis; it is an example of a disease mention of prostate cancer annotated as positive in Step 2-a but as negative in Step 2-b. The text of *DM* is not a disease mention in this study because diabetes mellitus (DM) is not the object disease. In Japan, the disease mentions in EHR are written in semi-structured problem lists based on the Problem Oriented Medical Record (POMR) format [28] with the # sign or narrative notes.

Table 2
Four clinical note datasets classified according to data types and formats in which disease mentions are written. The dataset types (Summary-st), (Summary-nr), (Progress-st), and (Progress-nr) are mutually exclusive.

Dataset type	Data type	Format
(Summary-st)	Discharge summaries	Semi-structured problem lists in discharge summaries (with the # sign in the same sentence as the disease mention)
(Summary-nr)		Narrative notes written in free-text space
(Progress-st)	Progress notes and other free texts	Semi-structured problem lists in progress notes (with the # sign in the same sentence as the disease mention)
(Progress-nr)		Narrative notes written in free-text space

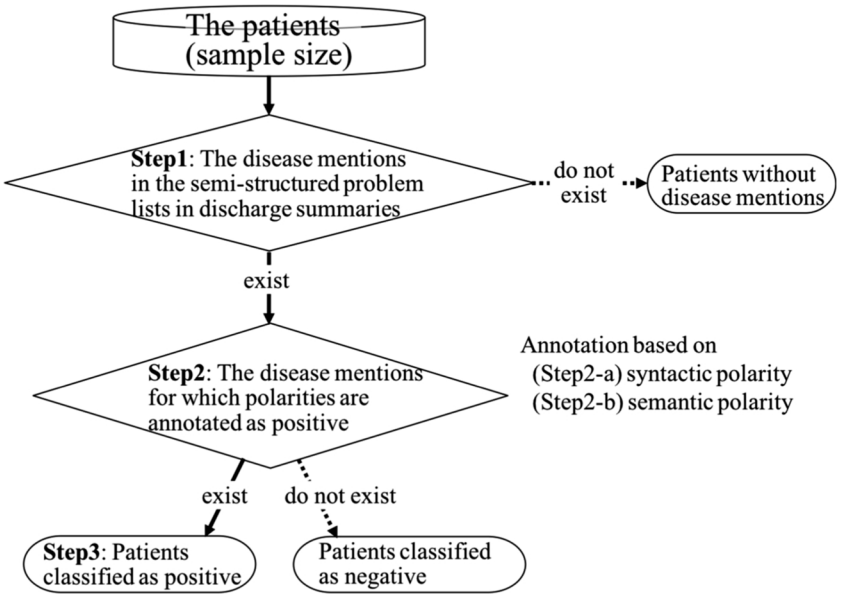

Fig. 2. Classification of the patients with an example using semi-structured problem lists in the discharge summaries (Table 2 (Summary-st)). This classification was repeated for each dataset type listed in Table 2 and for each object disease.

classified as semantic positive, respectively. It is important to note that this study focused on EHR-based phenotyping, which is for identifying patients rather than disease mentions. Therefore, we targeted number of patients rather than number of disease mentions.

2.2.2. Classifying physicians' intentions to write disease mentions annotated as syntactic polarity positive but semantic polarity negative

A set of exhaustive and mutually exclusive labels were created to capture the physicians' intentions to write disease mentions annotated as positive by syntactic polarity but negative by semantic polarity (Table 1 (B)). Similar to a previous study by Colicchio et al. [27], we conducted a modified Delphi process in which the first version of the classification was shared with the study's co-authors, who then provided suggestions iteratively until a consensus was reached (Appendix D). As shown below, we clarified the ratio of the number of patients whose clinical notes included the disease mentions in each category to the sample sizes for each disease. When one patient was classified into different categories, all categories were counted.

3. Results

3.1. Number of patients classified as positive based on the two polarity principles

Table 3 presents the proportion of patients respectively classified as syntactic positive and semantic positive to the sample size (shown in Table A. 1. in Appendix A) for 10 object diseases and four dataset types (Please refer to Appendix E). The differences between syntactic positive and semantic positive quantified the proportion of patients whose clinical notes include disease mentions that do not signify their diagnosis, even without any modifier or indicator for each disease and each dataset type. The ratios of patients classified as semantic positive to patients classified as syntactic positive (Table 3 < B > / < A > to < H > / < G >) are not 1.0 for more than two object diseases for each dataset type. The fact that the ratio is not 1.0 means that some disease mentions do not indicate the patients' diagnosis even without any modifier or indicator. For the free text in the progress notes, the ratio of patients classified as semantic positive to those classified as syntactic positive (Table 3 < H > / < G >) was only 78.1% on average (100.0% [DN] - 33.3% [HT]). This value could be interpreted as the influence on the performance of EHR-based phenotyping by extracting disease mentions from clinical notes (see details in Discussion).

3.2. Classifying disease mentions annotated as positive by syntactic polarity but negative by semantic polarity

The physicians' intentions to write the disease mentions (except the patients' diagnoses) were inductively classified into eight labels: differential diagnosis, misinterpretation of meaning, possibility of suffering from the disease in the future, screening, pre-surgery screening, general meanings, family history, and diagnosis of another person. General meanings do not pertain to the patients, but mean the disease itself. The screening, pre-surgery screening, and general meaning labels implied the same meaning as the generic modifier [22]; therefore, they were placed into the same category (Category 4). Family history and diagnosis of another person implied the same meaning as the subject modifier [22] and were placed into the same category (Category 5). In summary, the intentions to write disease mentions excluding the patient's diagnosis were classified into five categories. Category 1 was similar to the modifier "rule-out" used in MedLEE [29]. Category 3 was similar to the conditional modifier in Elhadad et al. [22]. Table 4 presents examples of disease mentions; the ratio of patients whose

Table 3
Ratio of the patients classified as positive based on the two principles of polarity to the sample size.

Patients classified	T2DM	HT	DN	PBC	crohn	prostate	breast	RCC	DAA	PE	average
(Summary-st): syntactic positive < A >	1.92%	0.00%	1.33%	1.38%	4.29%	3.00%	2.89%	3.16%	1.16%	2.00%	2.11%
(Summary-st): semantic positive < B >	1.92%	0.00%	1.33%	1.38%	4.29%	3.00%	2.77%	2.89%	1.16%	2.00%	2.07%
< B > / < A >	1.0		1.0	1.0	1.0	1.0	0.958	0.917	1.0	1.0	0.981
(Summary-nr): syntactic positive < C >	1.73%	0.00%	1.00%	5.17%	7.14%	9.25%	10.84%	9.47%	4.89%	4.86%	5.44%
(Summary-nr): semantic positive < D >	1.73%	0.00%	1.00%	4.48%	6.43%	8.50%	10.48%	8.16%	4.42%	4.86%	5.01%
< D > / < C >	1.0		1.0	0.867	0.900	0.919	0.967	0.861	0.905	1.0	0.921
(Progress-st): syntactic positive < E >	5.00%	0.00%	1.67%	6.20%	10.00%	14.50%	6.87%	11.05%	4.88%	6.00%	6.62%
(Progress-st): semantic positive < F >	5.00%	0.00%	1.67%	5.86%	9.29%	14.50%	6.75%	10.79%	4.65%	6.00%	6.45%
< F > / < E >	1.0		1.0	0.944	0.929	1.0	0.983	0.976	0.952	1.0	0.974
(Progress-nr): syntactic positive < G >	3.46%	0.46%	2.67%	12.76%	20.36%	23.50%	24.81%	20.53%	13.49%	14.28%	13.63%
(Progress-nr): semantic positive < H >	3.27%	0.15%	2.67%	8.62%	12.86%	21.50%	21.20%	15.79%	11.86%	8.57%	10.65%
< H > / < G >	0.944	0.333	1.0	0.676	0.610	0.916	0.854	0.769	0.879	0.600	0.781

clinical notes included the disease mentions classified in each category is shown in Table 5.

Category 1: Differential diagnosis

Category 2: Misinterpretation of meaning (Category 2 is not a typographic error [26].)

Category 3: Possibility of suffering from the disease in the future

Category 4: Screening, pre-surgery screening, general meanings

Category 5: Family history, diagnosis of another person

Table 5 shows that the disease mentions in the progress notes comprised more varieties of intentions than those in discharge summaries. Progress notes included disease mentions classified as Categories 3 or 4, which are not found in the discharge summaries. This suggests that the physicians' daily thought processes were described in the progress notes more than the discharge summaries. Therefore, it is difficult to create the annotated corpus including all clinical notes compared to those including only the discharge summaries.

4. Discussion

4.1. Difference between syntactic polarity and semantic polarity

Our manual review of 487,300 clinical notes for 10 diseases clarified the presence of disease mentions that do not connote the patient's diagnosis contrary to syntactic characteristics for all object diseases, except DN. This is the first large-scale manual review of clinical notes.

For DN, all patients classified as syntactic positive were also classified as semantic positive. This was because physicians wrote just *diabetic nephropathy* (no description of stage; it is not the object disease mentions) in clinical notes when physicians intended other meaning than the patient's diagnosis, while our definition of DN was over stage 2 DN. The smallest ratio of patients classified as semantic positive to patients classified as syntactic positive was 33.3% for HT (Table 3 < H > / < G >). This was because physicians wrote just *hypertension* in progress notes for patients suffering from essential hypertension. Patients whose clinical notes included disease mentions of HT suffered from refractory hypertension, and the physicians' intentions to describe disease mentions of HT were differential diagnosis (Table 5). This reflects the physicians' thought process for deciding the final diagnosis. We assumed why physicians wrote disease mentions, which do not connote the patient's diagnosis, without any modifier or indicator. It would be because physicians, who have background medical knowledge and tacit knowledge about clinical practices, can easily understand the intentions of the disease mentions regardless the existence of modifiers or indicators; therefore, in order to reduce the time-cost in writing descriptions of concepts physicians understand inherently, it is postulated they skipped writing modifiers or indicators

Table 4
Examples of physicians' intentions to write disease mentions that do not indicate that the patient has the disease. The disease mentions of the object diseases analyzed in this study are underlined.

Category	*Examples of clinical notes* (explanations of physicians' intentions to write disease mentions)
1: Differential diagnosis	*Prostate cancer* *Atypical adenomatous hyperplasia* *No tumor found.* (This disease mention indicates one of the differential diagnoses of the patient before the biopsy.)
2: Misinterpretation of meanings	*Crohn's disease good outcome* (The patient was diagnosed with ulcerative colitis. However, one physician, whose specialties did not include gastroenterology, misunderstood the patient's final diagnosis, and wrote *Crohn's disease* as the patient's diagnosis repeatedly. Ulcerative colitis and Crohn's disease are similar and easy to confuse – even for physicians. It was not a false diagnosis.)
3: Possibility of suffering from the disease in the future	*PE caution* (The patient rested on a bed after the operation and had an elevated risk for PE. This description implied that the physicians were cautious about PE, and the patient did not suffer from PE. However, there was no modifier, such as "will" or "if," in the same sentence. *Caution* is a one noun and not a verb. *PE caution* is a grammatically incorrect phrase as both *PE* and *caution* are nouns. However, there were similar grammatically incorrect phrases in Japanese clinical notes that listed nouns. When this phrase was rewritten to be a grammatically correct sentence, this phrase would be "Please beware of PE because the patient could suffer from it.")
4: Screening, pre-surgery screening, general meanings	*Breast cancer – last 2007* (There were no descriptions or data suggesting that the patient had breast cancer in her EHR. Breast cancer screening is recommended every other year for women over 40 years old in Japan, and this description implied that the patient's last breast cancer screening was in 2007.)
5: Family history, diagnosis of another person	*"My son was hospitalized from March 2014. (He) said that the physicians did not discover the disease. In April, (he) was found to have Crohn's disease. (He) is suffering from Crohn's disease, MR, and epilepsy."* (In "pro-drop" languages such as Japanese, pronouns are often unrealized in text; for example, the subject of *was* or *is* (the son of the patient) is omitted in these sentences. Please note that these sentences reflect the patient's own statement.)

Table 5
Ratio of patients whose clinical notes included disease mentions, classified in each category to the sample size. Hyphen means 0.00%.

Dataset type	CAtegory	T2DM	HT	DN	PBC	crohn	prostate	breast	RCC	DAA	PE
(Summary-st)	**1**	-	-	-	-	-	-	0.12%	0.26%	-	-
	2	-	-	-	-	-	-	-	-	-	-
	3	-	-	-	-	-	-	-	-	-	-
	4	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	-	-
(Summary-nr)	**1**	-	-	-	0.69%	0.36%	0.75%	0.36%	0.79%	-	-
	2	-	-	-	-	-	-	-	-	0.23%	-
	3	-	-	-	-	-	-	-	-	-	-
	4	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	0.36%	-	-	0.53%	0.23%	-
(Progress-st)	**1**	-	-	-	0.34%	0.71%	-	0.12%	0.26%	-	-
	2	-	-	-	-	-	-	-	-	-	-
	3	-	-	-	-	-	-	-	-	-	-
	4	-	-	-	-	-	-	-	-	-	-
	5	-	-	-	-	-	-	-	-	0.23%	-
(Progress-nr)	**1**	-	0.31%	-	1.72%	2.86%	1.25%	0.96%	4.21%	0.70%	1.43%
	2	-	-	-	1.03%	0.71%	-	0.12%	-	0.23%	-
	3	-	-	-	-	-	0.50%	0.96%	-	0.47%	3.14%
	4	-	-	-	1.38%	3.57%	0.50%	1.08%	0.26%	-	1.14%
	5	0.19%	-	-	-	0.71%	-	0.48%	0.26%	0.23%	-

(Appendix F).

The ratio of patients classified as semantic positive to those classified as syntactic positive (Table 3 < B > / < A > ˜ < H > / < G >) could be interpreted as the bias of the precision value if named entity recognition (NER) and factuality analysis (FA) are adopted as simple and robust EHR-based phenotyping algorithms to be shared across institutions or countries [16,17,30]. For example, when the object disease is breast cancer and the object clinical notes are only semi-structured problem lists in discharge summaries, the percentage of patients who have the disease is 95.8% of those with the disease mentions identified by NER and FA (which are identified based on syntactic characteristics). NER and FA are well-known NLP techniques and have actively been used on clinical notes in English [31,32], Japanese [33–35], and other languages [36–38]. In the context of EHR-based phenotyping, factuality is the certainty and polarity of whether a patient has an object disease; this is an expansion of the definition by Sauri and Pustejovsky [39]. When an EHR-based phenotyping researcher reproduces an existing technique or bio-medical researchers use an existing technique, the value of precision would not be reproduced and the comparison between the published results and the reproduced results would be meaningless unless the researchers unveil whether such disease mentions were regarded as positive or negative. To prevent such a problem, our classification of inaccurate disease mentions through clarifications of clinical experts' knowledge would enable accurate annotation even if annotators are not clinical professionals.

4.2. Future significance of the classification of the physicians' intentions to write disease mentions

Our next challenge is to distinguish disease mentions that cause bias from disease mentions that signify a patient's disease to decrease error in EHR-based phenotyping algorithms [16,17,30–38].

Category 5 includes the diagnosis of people other than the patient's family (such as the patient's neighbors), while only family history in Category 5 has been targeted directly to automatically distinguish disease mentions signifying patients' diagnoses [40]. Therefore, an automatic technique to distinguish disease mentions that connote another person's diagnosis from the patient's diagnosis is required. Categories 2 and 3 consist of only a few patients; therefore, automatic discrimination may be difficult. However, domain-specific feature engineering by people is effective for cases difficult to classify using machine learning [41,42]. Our categories could be regarded as domain-specific features engineered by physicians; they might be useful for automatically distinguishing disease mentions classified in Categories 2 or 3 from other intentions. Designing an EHR might be an effective solution for distinguishing Categories 1 or 4 from other intentions. For example, the EHR's structural description sections (which make physicians write about differential diagnosis or screening) might lead to reduced disease mentions classified in Categories 1 or 4 written in free texts. Indeed, we could not provide a new solution to distinguish these disease mentions that cause the bias from the disease mentions that signify the patient's disease. However, highlighting these new challenges to researchers in NLP or machine learning could be informative to wide areas of data analysis.

4.3. Limitations and future works

This pilot study was conducted based on the data obtained from one Japanese educational hospital. In Japan, narrative notes in EHR were written in Japanese, and unforeseen problems caused by this language might remain. Future studies will aim to solve the new challenges demonstrated by the classification of physicians' intentions.

5. Conclusion

This study clarified that the semantic polarity of a patient's diagnosis cannot be judged by only syntactic characteristics, and quantified that, on average, the ratio of patients classified as semantic positive to those classified as syntactic positive was only 78.1%. This value might indicate the influence of inaccurate disease mentions on the precision of EHR-based phenotyping by extracting disease mentions from clinical notes. We hope that our results based on each dataset will help researchers in diverse research environments with different available dataset.

Authors' contributions

RK, ES, YK, TI, and OK formulated the study concept and design; RK performed the data analysis; and all authors contributed to interpreting the results and writing the manuscript.

Funding

This work was supported by JSPS KAKENHI [grant numbers 16J05555, 16K09161]; and Research Grant for Lecturers in University of Tsukuba Hospital (2018).

Ethics

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, 10791 (2015)).

Statement on conflicts of interest

The authors have no competing interests to disclose.

Declarations of interest

None.

Summary table

"what was already known on the topic"

- EHR-based phenotyping cannot achieve high performance because clinical notes include disease mentions that do not signify patients' diagnoses.
- The ratio of disease mentions that do not signify patients' diagnoses to those that signify patients' diagnoses is not quantified for many diseases.
- Annotation guidelines of disease mentions, and modifiers or indicators (which show that the disease mention does not signify patient's diagnosis) have been developed.
- The disease mentions that signify family history have been targeted directly so they can be distinguished automatically from the disease mentions signifying patients' diagnosis.

"what this study added to our knowledge"

- For more than two object diseases for each dataset type, the presence of disease mentions do not signify a patient's diagnosis even without any modifier or indicator in the same sentences has been clarified.
- If named entity recognition and factuality analysis are adopted as simple and robust EHR-based phenotyping algorithms, the bias occurred owing to disease mentions that do not signify patients' diagnosis in the value of precision is 78.1% (on average) for free text in progress notes.
- The five categories for physicians' intention to write disease intentions (except patients' diagnosis) demonstrated the new challenges that must be overcome in order to distinguish disease mentions that connote such intentions from those that connote the patient's diagnosis.

Appendices A–F

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2018.12.004.

References

[1] W.R. Hersh, Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance, Am. J. Manag. Care 13 (2007) 277–278.
[2] C. Safran, et al., Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper, J. Am. Med. Inform. Assoc. 14 (2007) 1–9.
[3] MIT Critical Data, Secondary Analysis of Electronic Health Records, Springer, 2016.
[4] N. McCormick, et al., Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis, PLOS One 9 (2014) e104592, , https://doi.org/10.1371/journal.pone.0104519.
[5] 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, , https://doi.org/10.1371/journal.pone.0140533.
[6] M. Fury, et al., The Implications of Inaccuracy: Comparison of Coding in Heterotopic Ossification and Associated Trauma, Orthoped. 40 (2017) 237–241, https://doi.org/10.3928/01477447-20170208-02.
[7] J.C. Kirby, et al., PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inform. Assoc. 23 (2016) 1046–1052, https://doi.org/10.1093/jamia/ocv202.
[8] C. Shivade, A review of approaches to identifying patient phenotype cohorts using electronic health records, J. Am. Med. Inform. Assoc. 21 (2014) 221–230, https://doi.org/10.1136/amiajnl-2013-001935.
[9] C.M. Delude, et al., The details of disease, Nature 527 (2015) S14–S15, https://doi.org/10.1038/527S14a.
[10] J. Pathak, et al., Electronic health records-driven phenotyping: challenges, recent advances, and perspectives, J. Am. Med. Inform. Assoc. 20 (2013) e206–e211, https://doi.org/10.1136/amiajnl-2013-002428.
[11] G. Hripcsak, et al., Next-generation phenotyping of electronic health records, J. Am. Med. Inform. Assoc. 20 (2012) 117–121, https://doi.org/10.1136/amiajnl-2012-001145.
[12] R.L. Richesson, et al., A comparison of phenotype definitions for diabetes mellitus, J. Am. Med. Inform. Assoc. 20 (2013) e319–e326, https://doi.org/10.1136/amiajnl-2013-001952.
[13] X. Jie, et al., Review and evaluation of electronic health records-driven phenotype algorithm authoring tool for clinical and translational research, J. Am. Med. Inform. Assoc. 22 (2015) 1251–1260, https://doi.org/10.1093/jamia/ocv070.
[14] V. Papez, et al., Evaluation of Semantic Web Technologies for Storing Computable Definitions of Electronic Health Records Phenotyping Algorithms, AMIA Annu, Symp. Proc. 2017 (2017) 1352–1361.
[15] C. Kotfila, et al., A systematic comparison of feature space effects on disease classifier performance for phenotype identification of five diseases, J. Biomed. Inform. 58 (2015) S92–S102, https://doi.org/10.1016/j.jbi.2015.07.016.
[16] W.Q. Wei, et al., Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus, J. Am. Med. Inform. Assoc. 19 (2012) 219–224, https://doi.org/10.1136/amiajnl-2011-000597.
[17] K.M. Newton, et al., Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network, J. Am. Med. Inform. Assoc. 20 (2013) e147–e154, https://doi.org/10.1136/amiajnl-2012-000896.
[18] S. Gehrmann, et al., Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives, PLOS ONE 13 (2018) e0192360, , https://doi.org/10.1371/journal.pone.0192360.
[19] W. Chapman, et al., Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, J. Am. Med. Inform. Assoc. 18 (2011) 540–543, https://doi.org/10.1136/amiajnl-2011-000465.
[20] T.B. Murdoch, et al., The Inevitable Application of Big Data to Health Care, JAMA 13 (2013) 1351–1352, https://doi.org/10.1001/jama.2013.393.
[21] W.Q. Wei, et al., Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance, J. Am. Med. Inform. Assoc. e1 (2016) e20–e27, https://doi.org/10.1093/jamia/ocv130.
[22] N. Elhadad, et al., ShARe Guidelines for the Annotation of Modifiers for Disorders in Clinical Notes, (2018) (Accessed 19 Mar 2018), http://alt.qcri.org/semeval2015/task14/data/uploads/share_annotation_guidelines.pdf.
[23] R. Kagawa, et al., The impact of "possible patients" on phenotyping algorithms: Electronic phenotype algorithms can only be reproduced by sharing detailed annotation criteria, Stud. Health Technol. Inform. 245 (2017) 432–436.
[24] C. Holmes, The Problem List beyond Meaningful Use, Part 1, J. Am. Health Inform. Manag. Assoc. 81 (2011) 32–35.
[25] J.C. Krauss, et al., Is the problem list in the eye of the beholder? An exploration of consistency across physicians, J. Am. Med, Inform. Assoc. 23 (2016) 859–865, https://doi.org/10.1093/jamia/ocv211.
[26] NTCIR MedNLP Organizers, NTCIR11 MedNLP2 annotation Guidelines version 1.0, (2018) (Accessed 19 Mar 2018, http://mednlp.jp/ntcir11/guideline.pdf.
[27] T.K. Colicchio, et al., Health information technology adoption: Understanding research protocols and outcome measurements for IT interventions in health care, J. Biomed. Inform. 63 (2016) 33–44, https://doi.org/10.1016/j.jbi.2016.07.018.
[28] L.L. Weed, Medical Records That Guide and Teach, N. Eng. J. Med. 278 (1968) 593–600.
[29] C. Friedman, et al., Automated encoding of clinical documents based on natural language processing, J. Am. Med. Inform. Assoc. 11 (2004) 392–402, https://doi.org/10.1197/jamia.M1552.
[30] R. Kagawa, et al., Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach, J. Diabetes Sci. Technol. 11 (2017) 791–799, https://doi.org/10.1177/1932296816681584.
[31] G.K. Savova, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.
[32] H. Suominen, et al., Overview of the ShARe/CLEF eHealth Evaluation Lab 2013, International Conference of the Cross-Language Evaluation Forum for European Languages (2013) 212–231, https://doi.org/10.1136/jamia.2009.001560.
[33] E. Aramaki, et al., MedEx/J: A One-scan Simple and Fast NLP tool for Japanese Clinical Texts, Stud. Health Technol. Inform. 245 (2017) 285–288.
[34] E. Aramaki, et al., Overview of the NTCIR-11 MedNLP-2 Task, Proc. 11th NTCIR Conf. (2014), pp. 147–159.
[35] H. Imachi, et al., NTCIR-10 MedNLP Task Baseline System. Proc. 10th NTCIR Conf.

(2013) 710–712.
[36] N. Aurélie, et al., Clinical information extraction at the CLEF eHealth evaluation lab 2016, CEUR Workshop Proc. 1609 (2016) 28–42.
[37] Y. Xu, et al., Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries, J. Am. Med. Inform. Assoc. 21 (2014) e84–e92, https://doi.org/10.1136/amiajnl-2013-001806.
[38] A. Neveol, et al., Clinical Natural Language Processing in languages other than English- opportunities and challenges, J. Biomed. Seman. 9 (2018) 12–24, https://doi.org/10.1186/s13326-018-0179-8.
[39] R. Sauri, et al., Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text, Journal of Computational Linguistics 38 (2012) 261–299, https://doi.org/10.1162/COLI_a_00096.
[40] S. Goryachev, et al., Identification and Extraction of Family History Information from Clinical Reports, AMIA Annu. Symp. Proc. 2008 (2008) 247–251.
[41] P. Domingos, A Few Useful Things to Know about Machine Learning, Comm. ACM 55 (2012) 78–87, https://doi.org/10.1145/2347736.2347755.
[42] R. Takahama, et al., AdaFlock: Adaptive Feature Discovery for Human-in-the-loop Predictive Modeling, Proceedings of the 32nd AAAI Conf. Artif. Intell. (2018), pp. 1619–1626.