Detailed explanations of DA method

Detailed explanations of explicit information

Target disease: The four diseases we focus on in this study

Differential diagnoses: Guidelines are followed for each disease [15-18].

Upper disease: The ICD-10 tree structure is followed (Appandix Table S1). The full disease name of PBC is primary biliary chol angitis since approximately 2015 [14]. This study used the data until 2014; therefore, fibrosis and cirrhosis of liver are adopted a s the upper diseases of PBC.

Sibling diseases: The diseases which shares the upper disease an d has different cause with the target disease. Sister diseases wer e determined a priori.

Table S1–Definitions of exact and upper diseases

Target disease (UMLS CUI, ICD10 codes)	Upper Disease (ICD10 codes)
T2DM (C0011860, E11)	Diabetes Mellitus (E10-E14)
Essential HT (C0085580, I10)	Hypertensive diseases (I10-I15)
PBC (C0566602, K743)	Fibrosis and cirrhosis of liver (K74)
AIHA (C0002880, D591)	Acquired hemolytic anemia (D59)

The specific examples of differential diagnoses

HT: the disease names referred in the Indications/dosage and administration section of the attachment documents of anti-hypertensive, diuretic, β blocker, Ca blocker, or ARB/ACE inhibitor [ATC], except for hypertension (except for secondary hypertension) [JAPIC], OR the disease names referred as the specific situation which cause the hypertension under the specific situation in the guideline of hypertension.

PBC: the disease names referred in the Indications/dosage and administration section of the attachment document of Ursodeoxycholic acid [JAPIC]

The specific examples of sibling diseases

T2DM : type 1 DM, pregnancy diabetes, and other secondary DM [15]. Regarding overt diabetes in pregnancy, we decide whether T2DM or sister disease using with data after childbirt. Type 1 DM and secondary DM belong to the lower ICD-10 stage than DM(E10-E14).

HT: secondary hypertension, pregnancy hypertension [16].

PBC: Fibrosis and cirrhosis of liver caused by other diseases except for PBC (e.g. alcoholoc cirrhosis of liver, liver cancer).

AIHA: The disease (status) that can cause acquired hemolytic anemia except from AIHA (e.g. march hemoglobinuria, aortic valve replacement).

The detailed explanation of context annotation

(α) Increasing the conviction of the target disease

(i) The context which increase the conviction of the target disease.

(ii) The context which decrease the conviction of the differential diagnoses or siblings diseases.

(β) Probable upper disease AND no sibling diseases or differential diagnoses

There are context to increase the conviction of the upper disease AND no context to imply the sibling diseases or differential diagnosis

(γ) No other context

(δ) Possible diseases which are treated by the same medications used for the target disease

The patient is medicated with the same medication used for the target disease, but there are no context to imply to which diseases the drug is medicated. Moreover, there are no context to imply that the patient has the target disease and the patient has the disease, which could be treated by the same medications used for the target disease.

(ε) Decreasing the conviction of the target disease

(i) The definite descriptions of sibling diseases or differential diagnoses OR the context which increase the conviction of the target disease

(ii) There are definite description that the patient is medicated with the same medication for other diseases.

(ζ) Definite denial of the target disease

Detailed annotation criteria based on DA method

Each item of explicit information or context is independent, but some patients may have more than one item within explicit information or contexts. In such a case, annotate as follows;

(i) Annotators employ the newer explicit information or context in the following cases;

(1) The patient's explicit information or context changed following the results of lab tests or clinical examinations.

(2) The patient's explicit information or context changed following the changes of the patient's symptoms or medical conditions.

(ii) In other cases;

- If the patient has more than one item within explicit information, annotators employ the faster one in alphabet order, which means the higher conviction of the target disease; that is, when one patient has (a) and (d), annotators employ (a). If annotators judge the explicit information (a) is definitely incorrect, annotators would employ (d).)
- If the patient has explicit information (g) and any of (a)–(e), annotators employ (g) when the patient has context (ε) or (η). When the patient does not have context (ε) and (η), annotators employ the faster one in alphabet order in (a)–(e) (described above).
- If patient has context (α) and (β), annotators employ (α), which means the higher conviction of the target disease. If patient has context (δ)–(ζ), annotators employ the later one in Greek letters order; that is, when one patient has (δ) and (ε), annotators employ (ε).)
- If patient has more than one context of (δ)–(ζ), annotators employ the later one in Greek letters order.
- If patient has context (α) or (β), and any of context (δ)–(ζ),

  (i) when there are descriptions about coexistence of multiple diseases and the reason for the

judgement, annotators employ (α); that is, the patient has both T2DM and T1DM.

(ii) if the patient has any of explicit descriptions (a)–(e), annotators employ (α) or (β). If the patient has any of explicit descriptions (g), annotators employ the later one in Greek letters order among (δ)–(ζ).

The specific details of phenotyping algorithms

The specific details of GB phenotyping algorithms

Table S2– Phenotyping algorithms developed in this study

(i) T2DM: Patients with (A) AND ((B) OR (C)) modified eMERGE algorithm	
(A)	No DM-related ICD10 codes excluding T2DM-related codes (E10x and E12x and E13x), OR anti-GAD antibody ≦ 1.5 U/mL, OR anti-IA2 antibody ≦ 0.4 U/mL
(B)	Anti-diabetic medication
(C)	T2DM ICD-10 codes (E14x) AND abnormal lab test more than two times. Abnormal lab test is the following; HbA_{1c} ≧ 6.5 or HbA_{1c}(J) ≧ 6.1, OR glucose ≧ 126 mg/dL and (HbA_{1c} ≧ 6.5 or HbA_{1c}(J) ≧ 6.1), OR all glucose ≧ 126 mg/dL within one month (HbA_{1c} (J) was the HbA_{1c} standard used in Japan until March 2012, and HbA_{1c} (J) 6.1% is equivalent to HbA_{1c} 6.5%.)
(ii) Essential HT: Patients with (A) AND (B)	
(A)	No registration disease name with the phrase "secondary HT", "pregnancy HT," "hypertensive," "white coat HT," or some disease name that implies secondary HT (e.g., renal vascular HT)
(B)	Essential HT ICD-10 codes (I1x) except (A) OR Medication of ARB/ACE inhibitor
(iii) PBC: Patients with (A) AND ((B) OR (C))	
(A)	Anti-mitochondrial antibody (AMA) ≧ 20 times
(B)	(γGTP≧ 68 IU/L AND ALP≧ 359 IU/L) OR PBC ICD-10 codes (K743)
(iv) AIHA: Patients with ((A) AND (B)) OR (C)	
(A)	More than four abnormal labs as follows; Hemoglobin ≦ 13.8 g/dL (male) or 11.3 g/dL (female), Reticulocytes ≧ 2.0%, Haptoglobin ≦ 16 mg/dL, indirect Bilirubin ≧ 0.5 mg/dL, Urobilinogen in urine is positive, or hyperplasia of erythroblast cells in the bone marrow
(B)	Direct Coombs test is positive AND (AIHA ICD-10 codes (D591) OR no ICD10 codes, which mention the diseases that cases acquired hemolytic anemia (D50x-58x, D46x-48x, C81x-96x))
(C)	Disease names of AIHA in EHR, in applying the technique in [20]

The specific details of naïve phenotyping algorithms

T2DM：Patients with E11x (ICD-10 code)

HT: Patients with I1x (ICD-10 code)

PBC: Patients with K743 (ICD-10 code)

AIHA: Patients with D591 (ICD-10 code)

References

[ATC] KEGG, Anatomical Therapeutic chemical Classification, [cited 2016 November 3];
http://www.genome.jp/kegg-bin/get_htext?br08303.keg

[JAPIC] Japan Pharmaceutical Information Center, iyakuSearch [cited 2016 November 3];
http://database.japic.or.jp/is/top/index.jsp (in Japanese)