Detailed explanations of the DA method

Detailed explanations of explicit information

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

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

Upper disease: The *ICD-10* tree structure was followed (Appendix Table S1). The full disease name of PBC is “primary biliary cholangitis” since ca. 2015 [14]. This study used data until 2014; therefore, fibrosis and cirrhosis of liver are adopted as the upper diseases of PBC.

Sibling diseases: The diseases that share the upper disease and have causes different from those of the target disease. Sibling diseases were determined a priori.

Table S1–Definitions of exact and upper diseases

Target Disease (UMLS CUI, *ICD-10* codes)	**Upper Disease (*ICD-10* 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)

Specific examples of differential diagnoses

HT: The disease names referred in the Indications/dosage and administration section of the attachment documents of antihypertensive, diuretic, β-blocker, Ca blocker, or ARB/ACE inhibitor [appendix-1] medications, except for hypertension (except for secondary hypertension) [appendix-2], OR the disease names referred as the specific conditions that cause hypertension under the specific condition in the guidelines for hypertension.

PBC: The disease names referred in the Indications/dosage and administration section of the attachment document of ursodeoxycholic acid [appendix-2]

Specific examples of sibling diseases

T2DM: Type 1 DM, pregnancy diabetes, and other secondary DM [15]. Regarding overt diabetes in pregnancy, we decided whether it is T2DM or a sibling disease based on data obtained after childbirth. Type 1 and secondary DM belong to the *ICD-10* stage lower than that of DM (E10–E14).

HT: Secondary hypertension or pregnancy hypertension [16].

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

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

Detailed explanation of context annotation

(α) Increasing the conviction of the target disease

(i) The contexts that increase the conviction of the target disease.

(ii) The contexts that decrease the conviction of the differential diagnoses or sibling diseases.

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

Contexts to increase the conviction of the upper disease exists, but not context to imply the sibling diseases or differential diagnosis

(γ) No other contexts

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

The patient is medicated with the same medication used for the target disease, but no context exists to imply to which diseases the drug is medicated. Moreover, no context implies that the patient has the target disease and the patient has the disease, which could be treated with 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 contexts that increase the conviction of the target disease

(ii) Definite descriptions indicate that the patient is medicated with the same medication for other diseases.

(ζ) Definite denial of the target disease

Specific examples of explicit information or context items

Please refer to Table S2.

Detailed annotation criteria based on the DA method

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

(i) Annotators use the newer explicit information or context in the following cases:

(1) The patient’s explicit information or context changed after the results of lab tests or clinical examinations were obtained.

(2) The patient’s explicit information or context changed after the patient’s symptoms or medical conditions changed.

(ii) In other cases;

- If the patient has more than one item within an explicit information, annotators employ the faster one in alphabet order, which means higher conviction of the target disease. That is, when one patient has (a) and (d), annotators employ (a). If annotators judge that the explicit information (a) is definitely incorrect, annotators 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 contexts (ε) and (η), annotators employ the faster one in alphabetical order of (a)–(e) (described above).
- If patient has contexts (α) and (β), annotators employ (α), which means higher conviction of the target disease. If the patient has contexts (δ)–(ζ), annotators employ the latter in Greek-letter order; that is, when one patient has (δ) and (ε), annotators employ (ε).)

- If the patient has more than one of contexts (δ)–(ζ), annotators employ the latter one in Greek-letter order.
- If the patient has context (α) or (β), and any of contexts (δ)–(ζ),

  (i) when there are descriptions about the 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 the explicit descriptions (g), annotators employ the latter one in Greek-letter order among (δ)–(ζ).

Specific details of the phenotyping algorithms

Specific details of the GB phenotyping algorithms

Table S3– Phenotyping algorithms developed in this study

(i) T2DM: Patients with (A) AND ((B) OR (C)) modified eMERGE algorithm	
(A)	No DM-related *ICD-10* codes, excluding T2DM-related codes (E10x and E12x and E13x), OR anti-GAD antibody ≤1.5 U/mL, OR anti-IA2 antibody l ≤0.4 U/mL
(B)	Antidiabetic medication
(C)	T2DM *ICD-10* codes (E14x) AND abnormal lab test results more than two times. Abnormal lab test results for 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 1 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 names that imply secondary HT (e.g., renal vascular HT)
(B)	Essential HT *ICD-10* codes (I1x) except (A) OR medication with ARB/ACE inhibitor
(iii) PBC: Patients with (A) AND ((B) OR (C))	
(A)	Antimitochondrial 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 laboratory test results 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, and presence of urobilinogen in urine, or hyperplasia of erythroblast cells in the bone marrow
(B)	Positive direct Coombs test result AND [AIHA *ICD-10* codes (D591) OR no *ICD-10* codes, which mention the diseases that cases acquired hemolytic anemia (D50x-58x, D46x-48x, and C81x-96x)]
(C)	Disease names of AIHA in EHR in applying the technique in [20]

Specific details of the 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

[appendix-1] KEGG, Anatomical Therapeutic Chemical Classification, [cited 2016 November 3]; http://www.genome.jp/kegg-bin/get_htext?br08303.keg

[appendix-2] Japan Pharmaceutical Information Center, iyakuSearch [cited 2016 November 3]; http://database.japic.or.jp/is/top/index.jsp (in Japanese)

Table S2– Specific examples of items of explicit information or context. The italic font indicates the actual description in EHRs.

		T2DM	Essential HT	PBC	AIHA
Explicit information	(b)	—	—	*“There is the possibility of PBC or AIH”* *“questionable PBC”* *“The suspicion of asymptomatic PBC”*	*“There is the possibility of AIHA”*
	(e)	—	*“HT is likely.”*	—	—
Context	(α -i)	The patient is medicated with not insulin but antidiabetic medication. *“Diet therapy”*	*“There is no data which suggests secondary HT.”*	*“There is the possibility of the activity of PBC because of ALP/gGTP 400/100”*	
	(α -ii)	—	The values of lab tests, which need to deny the sibling diseases or differential diagnoses, are normal.	The values of lab tests, which need to deny the sibling diseases or differential diagnoses, are normal.	—
	(β)	The patient has abnormal glycol-albumin, and have not checked HbA1c. The patient has checked blood glucose or HbA1c over one year, at least once in three months.	The patient has descriptions of the value of blood pressure and the anti-hypertensive medications, and there are no descriptions which suggest diseases that are treated by the anti-hypertensive medications except for HT.	The results of abdominal ultrasonography or lab tests, which suggest liver cirrhosis.	The definite descriptions of the diseases, which are referred as the cause of secondary AIHA.
	(ε -i)	*“Pancreatic cancer”*	—	*“chronic hepatitis B”*	*“hemolytic anemia by mechanical valve”*
	(ε -ii)	—	“Arrhythmia disappears after the medications of anti-hypertensive medications.”	*“The patient has been medicated with ursodeoxycholic-acid and AST and ALT have become normalized.”*	—