A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease

Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, Anne-Sophie Jannot

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE

Open Access

A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease

Jean-Baptiste Escudié[1,2,3*], Bastien Rance[1,2], Georgia Malamut[1], Sherine Khater[1], Anita Burgun[1,2], Christophe Cellier[1] and Anne-Sophie Jannot[1,2]

Abstract

Background: Data collected in EHRs have been widely used to identifying specific conditions; however there is still a need for methods to define comorbidities and sources to identify comorbidities burden. We propose an approach to assess comorbidities burden for a specific disease using the literature and EHR data sources in the case of autoimmune diseases in celiac disease (CD).

Methods: We generated a restricted set of comorbidities using the literature (via the MeSH® co-occurrence file). We extracted the 15 most co-occurring autoimmune diseases of the CD. We used mappings of the comorbidities to EHR terminologies: ICD-10 (billing codes), ATC (drugs) and UMLS (clinical reports). Finally, we extracted the concepts from the different data sources. We evaluated our approach using the correlation between prevalence estimates in our cohort and co-occurrence ranking in the literature.

Results: We retrieved the comorbidities for 741 patients with CD. 18.1% of patients had at least one of the 15 studied autoimmune disorders. Overall, 79.3% of the mapped concepts were detected only in text, 5.3% only in ICD codes and/or drugs prescriptions, and 15.4% could be found in both sources. Prevalence in our cohort were correlated with literature (Spearman's coefficient 0.789, $p = 0.0005$). The three most prevalent comorbidities were thyroiditis 12.6% (95% CI 10.1–14.9), type 1 diabetes 2.3% (95% CI 1.2–3.4) and dermatitis herpetiformis 2.0% (95% CI 1.0–3.0).

Conclusion: We introduced a process that leveraged the MeSH terminology to identify relevant autoimmune comorbidities of the CD and several data sources from EHRs to phenotype a large population of CD patients. We achieved prevalence estimates comparable to the literature.

Keywords: Autoimmune diseases, Celiac disease, Electronic health records, Icd 10, Phenotype, Prevalence study, Diabetes mellitus, type 1, Dermatitis herpetiformis, Thyroiditis, autoimmune, Arthritis, rheumatoid, Lupus erythematosus, systemic, Multiple sclerosis, Sjogren's syndrome, Addison disease, Arthritis, juvenile, Hepatitis, autoimmune, Graves' disease, Myasthenia gravis, Polyendocrinopathies, autoimmune, Antiphospholipid syndrome

* Correspondence: jean-baptiste.escudie@aphp.fr
[1] Georges Pompidou European Hospital (HEGP), AP-HP, Paris, France
[2] INSERM UMRS 1138, Paris Descartes University, Paris, France
Full list of author information is available at the end of the article

Background

Precise phenotyping of patient data remains one of the key points of personalized medicine. From a clinical perspective, detailed knowledge of the comorbidities enables targeted treatment strategies. And from a research perspective, specific and accurate phenotypes allow the recruitment of patients in observational or interventional studies. Clinical Data Warehouses (CDW) gather information on hundred thousands of patients. These CDWs can be used to phenotype comorbidities as in our institution [1].

To phenotype patients using Electronic Health Records (EHRs), many different EHR sources could be mined: for instance International Classification of Diseases (ICD) codes, clinical reports, drug prescriptions, procedures, and laboratory test results if relevant. ICD codes have been widely used to phenotype patients [2]. Several studies [3, 4] showed that billing codes were specific enough to identify patients suffering from a given condition such as inflammatory bowel disease. However, ICD codes have a low sensitivity, particularly if the main claim was another condition, because then the studied condition is less likely to be coded. Therefore, ICD codes might perform poorly to phenotype comorbidities. The clinical narrative within the EHRs forming the patient's medical history contains lots of detailed information such as data collected outside the hospital, e.g., results of lab tests performed before the admission, or information regarding decision support, e.g., rejected clinical hypotheses. Disorder terms are indeed present in various types of clinical narratives, such as radiology reports [5, 6], medical observations [7, 8], nurse narratives [9]– and generally every document produced for healthcare activity. Several authors reported that clinical notes offer good sensitivity; moreover combining billing codes, clinical notes, and medications provides superior phenotyping performance [2, 10–12].Strategies such as phenotyping algorithms combining different types and sources of data (e.g. as proposed by PheKB [13]) are promising, but they require expert time to develop and the number of available algorithms are limited and focused on specific diagnoses.

Celiac disease (CD) is an autoimmune disorder triggered by gluten in genetically susceptible individuals. The disease is characterized by autoantibodies directed against gluten such as anti-gliadin or other targets (anti-transglutaminase, anti-endomysial). Many symptoms can be associated with CD, the most prominent are caused by malabsorption. CD is also associated with numerous autoimmune comorbidities. These comorbidities add to the high burden of symptoms and complications for these patients, and might be target for specific treatment strategies and screening programs. This is why it is necessary to identify these subpopulations for clinical research and public health policies. Previous epidemiological studies have shown that dysthyroidism and type 1 diabetes mellitus were the most prevalent diseases in patients with CD (6.0 to 30.2% and 2.2 to 6.5% respectively) [14–20]. Several methods were used, namely autoantibody detection [17–20], questionnaires [14, 15, 21, 22], national register [23] and retrospective reviews of medical records [16, 24], providing heterogeneous results, and only estimates for one or few autoimmune comorbidities. The estimated prevalence of thyroid disorders varies largely in these studies (from 6 to 30%). Other authors have studied the prevalence of CD among patients with autoimmune diseases [25]. For example, there is a high prevalence of CD in young diabetic patients [26].

To the best of our knowledge, there is no clear review on the most prevalent set of autoimmune comorbidities associated to CD, while there is a need to phenotype autoimmune comorbidities burden in CD patients to enable further stratification of patients' profiles. While methods to phenotype patients for a specific disease exists, they do not allow to estimate comorbidities burden in a specific domain. This step requires usually domain experts to define a set of specific comorbidities, but this elicitation step might introduce bias. Biomedical literature and its metadata can also be mined to extract knowledge for various purposes [27]: precision medicine and drug repositioning, pathway extraction, gene function prediction, data integration, pharmacogenomics, toxicology. As the biomedical literature also explores comorbidities associated to diseases, its metadata could provide information on relevant comorbidities for a given disease.

The present study aimed to show how a workflow based on both literature and EHRs information can help identifying relevant autoimmune comorbidities in CD and to phenotype for these autoimmune comorbidities the population of adult CD patients followed up in Georges Pompidou European Hospital (HEGP). We assessed its performance in this context by assessing quantitatively whether literature-based knowledge was correlated to knowledge extracted from EHRs regarding autoimmune CD comorbidities. Our secondary objective aimed at assessing to what extent three major EHR components (ICD codes, drug prescriptions and narrative reports) contributed to identifying comorbidities in the specific domain of autoimmune comorbidities in our population.

Methods

Overview

We first selected the list of autoimmune diseases from MeSH® terminology. We then restricted this list to the most frequent autoimmune diseases associated with CD in the literature, based on the number of co-occurrences of MeSH® terms in MEDLINE®. We mapped selected autoimmune disorders to different terminologies to identify concepts and terms for data collection. Then we included CD patients from a local registry and completed by querying the hospital clinical data warehouse (CDW). Finally we identified status for selected autoimmune diseases for these patients by reviewing their EHR for mapped concepts.

Selection of a restricted set of autoimmune diseases using MeSH® co-occurrence file

To define autoimmune comorbidities burden in CD, we first extract a specific set of autoimmune disease. An initial list of autoimmune diseases was defined as: all the children of concept 'Autoimmune Disease' (*D001327*) in the MeSH® hierarchy, i.e., all the descriptors with a MeSH® *Tree Number* starting with *C20.111*. We identified amongst this list the autoimmune diseases most frequently associated with CD in MEDLINE® using the MeSH® co-occurrence file (MRCOC) provided by the U.S. National Library of Medicine as part of the UMLS Metathesaurus (2014AB release). The MRCOC file contains the number of times each pair of MeSH® descriptors occurred in MEDLINE® citations. We extracted all pairs of MeSH terms that contained both the MeSH descriptor for *Celiac Disease (D002446)* and any MeSH descriptor for autoimmune diseases. We arbitrarily restricted the list for this study to the 15 autoimmune diseases the most co-occurring with CD. This process allows to focus on a domain of comorbidities using the MeSH hierarchy and to a subset of comorbidities using the number of co-occurrences in MEDLINE®.

MeSH® mapping to drugs and diagnoses terminologies

We used the Unified Medical Language System (UMLS) to map autoimmune disorders MeSH® concepts to other terminologies used in the CDW. To leverage diagnosis codes we used the International Classification of Diseases version 10 (ICD-10). We used the Anatomical Therapeutic Chemical Classification System (ATC) whenever a drug was specific to an autoimmune disease. These terminologies (MeSH, ICD-10 and ATC) also provided terms to constitute a catalog of words for helping textual data review.

Study population

The HEGP hospital has a specialized consultation for CD patients with about 50 new patients recruited per year in the last decade. The gastroenterology department has maintained a list of patients with CD for research purposes. The study has been approved and data access granted by the following commissions: the *"Commission Nationale de l'Informatique et des Libertés"* (declaration #174350) and the *HEGP institutional review board* (registration #00001072).

To complete the list of patients maintained by the gastroenterology department, we also extracted patients fulfilling the three following criteria: (i) at least one encounter in the 2000-2014 period of time (ii) presence of at least one ICD 10 code for CD (K90) in billing claims; (ii) one or more hospitalization stay or consultation in the gastroenterology department and (iii) at least one text document (discharge or letter) containing the term 'celiac disease' or one of its synonyms. We extracted these data from HEGP i2b2 CDW [25]. This CDW contains routine care data divided into several categories among them demographics (age, sex, and vital status), vital signs (e.g., temperature, blood pressure, weight…), diagnoses (ICD-10), procedures (French classification), clinical data (structured questionnaires from EHR), free text reports, pathology codes (French ADICAP classification), biological test results (LOINC), and Computerized Provider Order Entry (ATC) drug prescriptions [28].

The study population counted 741 patients and a corpus of 6340 clinical reports (patients' inclusion flowchart available in Fig. 1).

Autoimmune diseases phenotyping

We queried the CDW to identify for each patient the presence/absence of the selected autoimmune diseases. We used the selected ICD-10 diagnosis codes to query billing data and selected ATC codes to query Computerized Prescription Order Entry data.

We extracted every narrative report including discharge summaries, outpatient reports, multi-disciplinary expert meeting summaries and letters from all departments of the hospital and reviewed them manually to extract selected autoimmune disease status for each patient. To facilitate the manual review, we developed a browser-accessible software, FASTVISU [29], interfaced with the CDW to display and explore all the documents for a given patient. FASTVISU highlights terms with an entity recognition module based on a list of regular expressions (e.g., the pattern \bdiab\w + \b to match for diabetes) defined by the user and approximate matching techniques. For the 15 selected autoimmune comorbidities, a set of regular expressions was established, and used to query the CDW and obtain the CD corpus; then two trained physicians reviewed all the clinical narratives in the CD corpus using FASTVISU to validate the presence/absence of each of the selected autoimmune diseases for each patient.

Because autoimmune diseases are chronic diseases, a patient was considered having the autoimmune disorder if the condition was ascertained at least once in her/his EHR. The analysis was performed considering all sources together then each source separately.

Figure 2 illustrates our workflow.

Statistical analysis

Patients' characteristics were summarized using median and interquartile ranges for quantitative variables and proportions for qualitative variables. We measured Sperrin's *I*

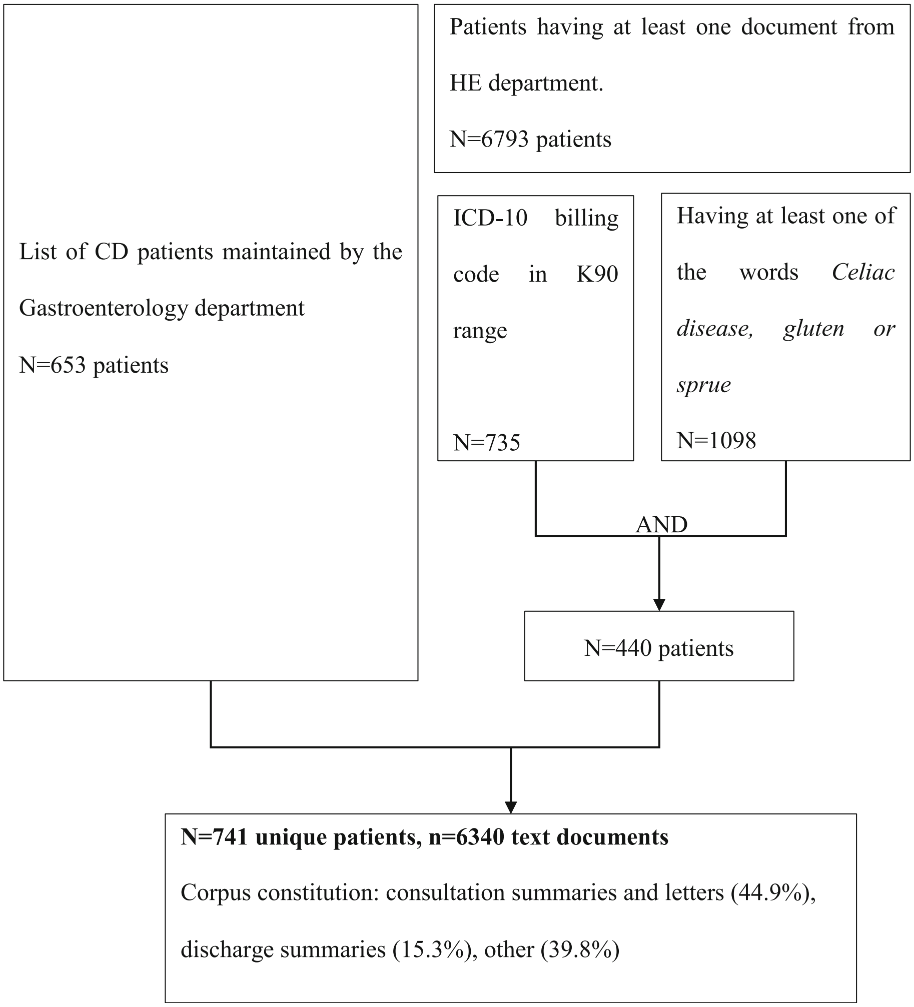
List of CD patients maintained by the Gastroenterology department
N=653 patients
Patients having at least one document from HE department.
N=6793 patients
ICD-10 billing code in K90 range
N=735
Having at least one of the words Celiac disease, gluten or sprue
N=1098
AND
N=440 patients
N=741 unique patients, n=6340 text documents
Corpus constitution: consultation summaries and letters (44.9%), discharge summaries (15.3%), other (39.8%)

Fig. 1 Patients inclusion flowchart

coefficient [30] to evaluate temporal irregularity of the data recorded in the EHR, denoting whether some encounters would provide more data and, consequently, whether other time periods would be less covered. Sperrin's I coefficient was calculated for each patient with at least 2 encounters using formula (1):

$$I = \frac{2}{n} + \frac{n-2}{n}\left[1-\sqrt{(n-1)\ Var\left(g_i;\ i = 1,\ldots,\ n-1\right)}\right] \qquad (1)$$

Reviewers' mutual-agreement in the text review was evaluated with Cohen's alpha coefficient.

For each autoimmune disease, a case was defined by at least one ICD code, one drug prescription, or one mention in text. For each prevalence estimate, we computed the 95% confidence interval or Wilson's interval if the proportion was close to 0%.

To compare the contributions of text, ICD codes and drugs to estimate autoimmune comorbidities, we computed the proportions of cases detected by text alone, structured information (ICD code or drug prescription on CPOE) alone, or both.

Finally, to assess the performance of literature-based comorbidities selection, we estimated the correlation between MeSH® and diseases prevalence, using Spearman's correlation coefficient between number of publications indexed with MeSH® terms for both CD and an autoimmune disease, and the prevalence the corresponded disease obtained from our EHR extraction.

Statistical analysis was conducted with R (version 3.1.2).

Results

We selected the first fifteen autoimmune disease whose MeSH® terms co-occurred the most with CD MeSH® terms in MEDLINE (Table 1). The most frequent were type I diabetes (523 citations), dermatitis herpetiformis

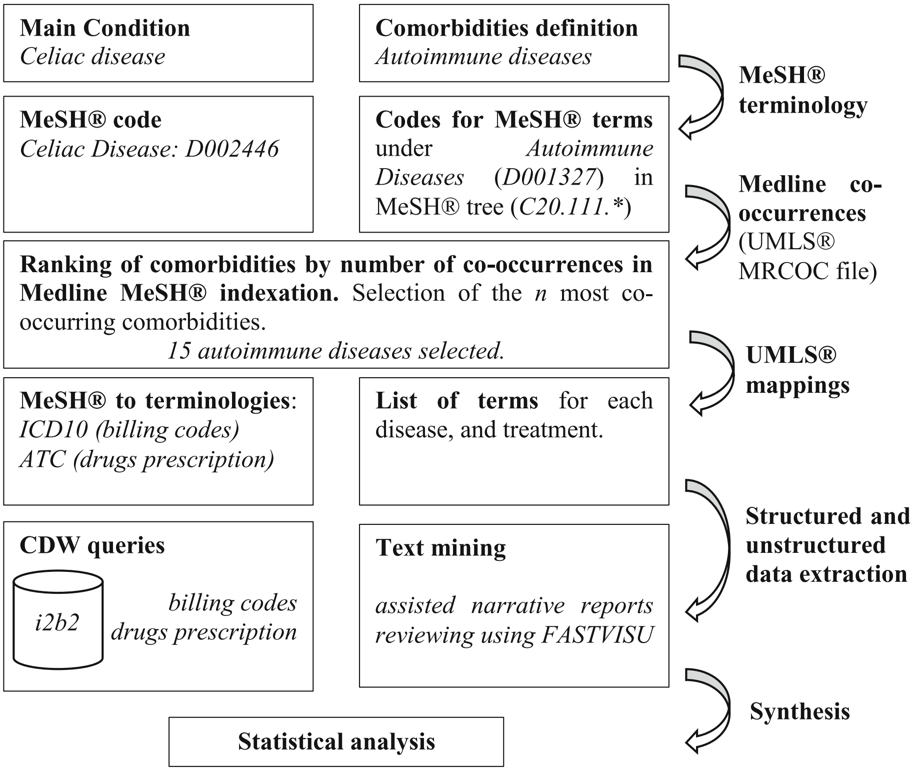
Main Condition
Celiac disease
Comorbidities definition
Autoimmune diseases
MeSH® terminology
MeSH® code
Celiac Disease: D002446
Codes for MeSH® terms under *Autoimmune Diseases* (*D001327*) in MeSH® tree (*C20.111.**)
Medline co-occurrences (UMLS® MRCOC file)
Ranking of comorbidities by number of co-occurrences in Medline MeSH® indexation. Selection of the *n* most co-occurring comorbidities.
15 autoimmune diseases selected.
UMLS® mappings
MeSH® to terminologies:
ICD10 (billing codes)
ATC (drugs prescription)
List of terms for each disease, and treatment.
Structured and unstructured data extraction
CDW queries
i2b2
billing codes
drugs prescription
Text mining
assisted narrative reports reviewing using FASTVISU
Synthesis
Statistical analysis

Fig. 2 Workflow from comorbidities selection to comorbidities burden phenotyping

(478 citations), and thyroiditis (96 citations). The number of co-occurrences ranged from 12 to 523.

The mapping of the selected autoimmune diseases to ICD-10 provided 39 diagnosis codes (Table 1). For the mapping of the selected disease to ATC, levothyroxine was used as marker for dysthyroidism, and insulin as a marker for type I diabetes. Insulin and levothyroxine corresponded to a total of 263 ATC codes (Additional file 1). We built a catalog of 55 terms from ICD-10, ATC and MeSH to help review of narratives.

The 741 included patients had overall 6340 clinical reports. Patients' characteristics are presented in Table 2. Ages spanned adulthood, with a mean of 42.5 years and a standard deviation of 16.9 years. Most were female, with a sex ratio of 2.8.Eighteen patients (2.4%) had only one encounter. One third of the patients had been followed up for 5 to 20 years. Patients had a median of 5 [3; 10] clinical documents, with a maximum of 146. Half of the CD patients had no other ICD-10 code than CD, 19.7% had 1 to 5 distinct ICD-10 codes, and 12.7% had between 6 and 69 codes. Most patients (93.5%) had at least one hospitalization stay during the 2000-2014 period, as all patients are offered a day hospital admission for initial management of their CD.

Sperrin's *I* coefficient was 0.759 mean (SD) 0.10, indicating that patients were followed up regularly.

Readers voted on 466 items. More specifically, 465 patients out of the 741 included patients had no highlighted terms and 466 autoimmune disease items had at least one highlighted term occurrence on the 276 remaining patients. For 140 items, voters both approved that the patient suffered from this disease; for 304 items, readers both disapproved that the patient suffered from this disease and for 22 items, readers mutually disagreed. Therefore, inter-reviewer agreement for autoimmune disorder identification in narrative reports was excellent, with a Cohen's kappa value of 0.89.

The prevalence estimates of the 15 selected autoimmune diseases with their 95% confidence intervals are presented in Tables 3 and 4 together with literature estimates.

Figure 3 represents the respective contributions of text and structured information to identify cases. Overall, 79.3% of the cases were detected only in text, 5.3% only in ICD codes and/or drugs prescriptions, and 15.4% could be found in both types of sources, with variations across the diseases. 86% of dermatitis herpetiformis diagnoses were mentioned only in narrative reports, whereas only 7% were found as structured information exclusively; in 7% of the cases the dermatitis herpetiformis information was present in both text and ICD codes. Information regarding autoimmune thyroiditis was mostly present only in text (92.5% of detected cases),

Table 1 Autoimmune disease selection by descending order of co-occurrence frequencies and ICD-10 codes used for diagnosis extraction

MeSH® terms	N co-occurrences	DUI*	ICD-10
Diabetes Mellitus, Type 1	523	D003922	E10.[a]
Dermatitis Herpetiformis	478	D003874	L130
Thyroiditis, Autoimmune	96	D013967	E063
Arthritis, Rheumatoid	87	D001172	M069.[a]
Lupus Erythematosus, Systemic	73	D008180	M32.[a]
Multiple Sclerosis	44	D009103	G35
Sjogren's Syndrome	43	D012859	M350
Addison Disease	42	D000224	E271|E272
Arthritis, Juvenile	37	D001171	M089.[a]
Hepatitis, Autoimmune	35	D019693	K754
Graves' Disease	30	D006111	E050|E05.[a]
Glomerulonephritis, IGA	27	D005922	N0330|N0170
Myasthenia Gravis	22	D009157	G700
Polyendocrinopathies, Autoimmune	15	D016884	E31.[a]
Antiphospholipid Syndrome	12	D016736	D686.[a]

* *DUI* Descriptor Unique Identifiers

[a]Means all descending nodes, | means OR

with only 2.2% of cases detected in codes alone, and 5.3% in both text and structured data (ICD code and/or drugs prescription). Type 1 diabetes was detected by text alone in 17.6% of the cases, codes alone in 23.5%, and found in both in 58.8% of the cases.

The three most prevalent diseases were thyroiditis (12.6%), type 1 diabetes (2.3%) and dermatitis herpetiformis (2.0%). 18.1% of CD patients had at least one of the 15 autoimmune diseases, in accordance with literature

Table 2 Population characteristics

Age in years, mean (SD)	42.5 (16.9)
Sex, n (%)	
Female	545 (73.6)
Male	196 (26.4)
Follow-up time in years per patient, n (%)	
0 (1 encounter)	18 (2.4)
0–1	248 (33.5)
1–2	91 (12.3)
2–5	134 (18.1)
5–20	250 (33.7)
In- or Outpatient, n (%)	
Outpatients	48 (6.5)
Inpatients	693 (93.5)
Number of encounters, median (IQR), max	3 (2, 4), 47
Documents per patient, median (IQR), max	5 (3, 10), 146
Sperrin's I irregularity indicator, mean (SD), *n* = 638	0.759 (0.10)
Number of distinct ICD-10 codes per patient, n (%) [a]	
0	372 (50.2)
1	129 (17.4)
2–5	146 (19.7)
6–69	94 (12.7)

IQR interquartile range

[a]Except code K900 for Celiac disease

Table 3 Prevalence estimates

Disease	N cases	Prevalence per 1000 patient [CI95]
Autoimmune thyroiditis	93	125.5 (101; 149)
Type 1 diabetes mellitus	17	22.9 (12.1; 33.5)
Dermatitis herpetiformis	15	20.2 (10; 30.2)
Rheumatoid arthritis	7	9.4 (2.5; 16.3)
Autoimmune hepatitis	7	9.4 (2.5; 16.3)
Graves' disease	6	8.1 (1.6; 14.5)
Sjogren's syndrome	6	8.1 (1.6; 14.5)
Addison disease	3	4 (1.4; 11.8)
Systemic lupus erythematosus	2	2.7 (0.7; 9.8)[a]
Juvenile arthritis	1	1.3 (0.1; 7.6)[a]
Multiple sclerosis	1	1.3 (0.1; 7.6)[a]
Autoimmune Polyendocrinopathies	1	1.3 (0.1; 7.6)[a]
Antiphospholipid syndrome	0	0 (0; 5.2)[a]
Myasthenia gravis	0	0 (0; 5.2)[a]

CI95: 95% confidence interval

[a]Wilson's intervals

Table 4 Comparison with prevalence in the literature

	This study	Cosnes et al. [33]	Iqbal et al. [2]	Volta et al. [5]	Collin et al. [22]	Størdal et al. [6]	Diamanti et al. [1]	Van der Pals et al. [13]	Sategna-Guidetti et al. [27]	Counsell et al. [34]
Study characteristics										
Population	741 adults	378 children, 546 adults	356 adults (> 12 yo)	770 adults	335 adults	3006 children	558 children	335 children	241 adults (untreated)	107 adults
Information source	EHR	questionnaire	questionnaire	questionnaire	hospital records	national register	serology	serology	serology	serology
Origin	France	France	Canada	Italy	Finland	Norway	Italy	Sweden	Italy	Scotland
Year	2015	2008	2013	2012	1994	2011	2011	2014	2001	1994
Prevalence, % (95% CI)										
Autoimmune thyroiditis	12.6 (10.1, 14.9)	6.0	10.6	26.3	7.5	1.4	12.0	7.7	30.2	15.0
Type 1 diabetes mellitus	2.3 (1.2, 3.4)	6.5	2.2	3	5.4	4.7				
Dermatitis herpetiformis	2.0 (1.0, 3.0)	3.1	13.5	4						
Rheumatoid disease	0.9 (0.3, 1.6)	0.7	4.5		1.8					
Autoimmune hepatitis	0.9 (0.3, 1.6)	1.2	0.0							
Sjögren's syndrome	0.8 (0.2, 1.5)	0.2	0.8		3.3					
Addison's disease	0.4 (0.1, 1.2)	0.2			0.6					
Systemic lupus	0.3 (0.1, 1.0) [a]	0.2	1.1							
Multiple sclerosis	0.1 (0.0, 0.8) [a]	0.1								
Antiphospholipid syndrome	0 (0.0, 0.5) [a]	0.2								
Myasthenia gravis	0 (0.0, 0.5) [a]	0.2								

CI95: 95% confidence interval, [a] Wilson's interval

estimates. Prevalence estimates strongly correlate with co-occurrence ranking for the 15 autoimmune diseases, as measured by a Spearman's correlation coefficient value of 0.789 (*P* value = 0.0005). The three most prevalent autoimmune diseases in our adult CD population appeared as the top three in the co-occurrence ranking.

Discussion

We successfully identify major CD autoimmune comorbidities using a novel data-driven workflow leveraging MeSH® terminology and Medline MeSH® co-occurrences. We mapped these comorbidities to terminologies used in EHRs to phenotype a new set of 741 patients achieving prevalence estimates comparable to the literature. One finding is the importance of clinical text reports as the most informative data source to phenotype patients was clinical narratives.

External validation of prevalence estimates

The mining of EHR data allowed us to include 741 patients, one of the largest population of adult CD patients used to report autoimmune diseases prevalence in CD patients to the best of our knowledge. In this hospital based study, the prevalence of autoimmune comorbidity was 18.1% (95% CI 15.4 –21.0). The three most prevalent autoimmune comorbidities were thyroiditis with a prevalence of 12.6% (95% CI 10.1–14.9), type 1 diabetes mellitus with 2.28% (1.2 –3.4) and dermatitis herpetiformis with 2.0% (95% CI 1.0–3.0).

We compared our prevalence estimates with literature as an external validation. Our disease prevalence estimates are in the highest range compared to other published studies [14–20, 23]. The first explanation is that we assessed prevalence from a hospital-based cohort from a CD specialized center, while most non-complicated CD disease, therefore with no additional autoimmune burden,

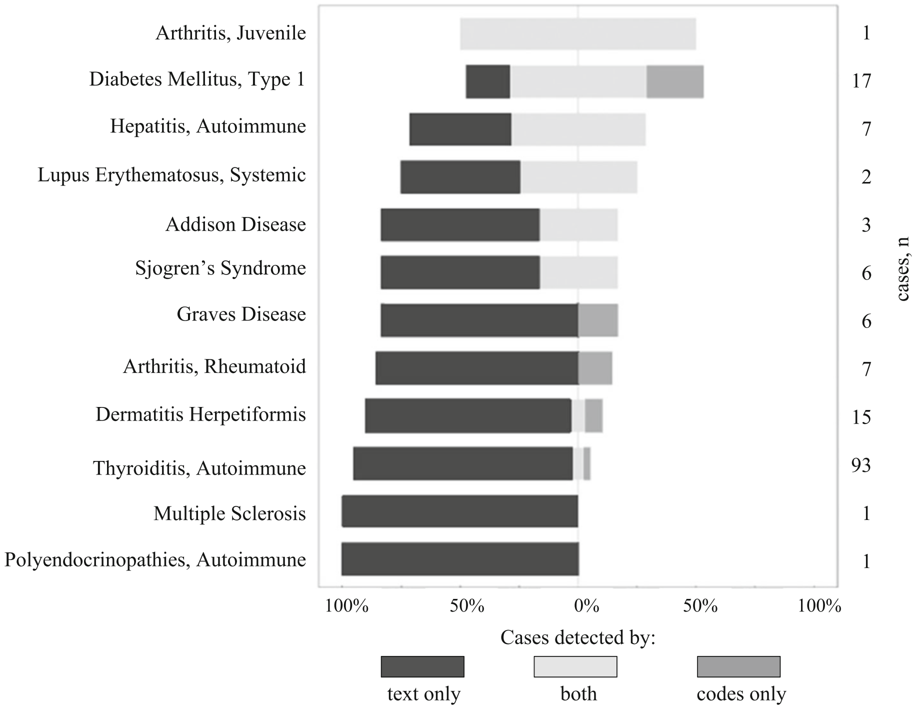

Fig. 3 Phenotypes identified by text reviewing only (black), ICD codes or drugs only (grey), both (light grey), in percent

are likely to be followed-up in ambulatory care only. Moreover, our study benefited from the coverage of the CDW, which includes a long follow-up and, therefore, increases the probability of mentioning an associated autoimmune disease. The quality of this longitudinal source of information was measured by Sperrin's coefficient, which demonstrates a broad coverage of text documents during the follow-up period. In contrast, prevalence studies based on questionnaires [14, 15, 21] may underestimate prevalence, e.g., due to memory bias.

It would have been of interest to extract diagnosis date, but as we expected many missing data and approximations due to early childhood diagnoses, we did not extract this information.

Text reports were more sensitive than ICD codes and medications for detecting autoimmune comorbidities

In this study, most diagnoses were ascertained through text reports. This finding is consistent with the review by Shivade et al. of 97 studies using EHR for phenotype identification [2]. A typical example is thyroiditis, where about 98% of the cases were found in the text reports (92.5% only in text, and 5.3% in both text and structured data).

Few diagnoses cases were ascertained through ICD-10 codes. In France, as in many countries, ICD-10 coding is primarily used for billing purposes and limited to inpatients. Consequently, the coding does not aim to cover all the patient's conditions [31]. Moreover narrative reports include extensive information such as a dedicated *medical history* section [10]. Additionally, autoimmune disease cases were identified in our study using documents produced during outpatient visits during which no ICD-10 codes were collected in line with French regulation (no ICD-10 codes are produced during outpatient visits in France). Similarly, Wei et al. analyzed the respective contributions of clinical notes, ICD codes and medications for detecting ten diseases in EHRs and showed that clinical narratives offered the best sensitivity (0.77) [12].

Our results showed the benefits of combining text mining and structured data extraction. Other examples are found in the literature in colon cancer [11], atrial fibrillation, Alzheimer's disease, breast cancer, gout, human immunodeficiency virus infection, multiple sclerosis, Parkinson's disease, rheumatoid arthritis, and types 1 and 2 diabetes mellitus [12].

Literature-based selection of autoimmune diseases

To the best of your knowledge there was no clear synthesis of major CD autoimmune comorbidities. The novel approach mining literature presented in this study enabled to identify relevant comorbidities as attested by the fact that the attention from the literature was coherent with the prevalence found both in the literature and in our cohort: autoimmune dysthyroidism or type 1 diabetes mellitus appeared in the top co-occurring MeSH® terms and these comorbidities were described in the literature as being the most prevalent autoimmune comorbidities in CD patients [16, 20]. Furthermore, the three most prevalent autoimmune diseases in our adult CD population appeared as the top three in the co-occurrence ranking. Our method is flexible as domain restriction using MeSH® hierarchy and limiting the number of results with the number of co-occurrence are both optional, although we haven't evaluated this method without these two types of restrictions. Moreover, this novel approach provided us with the most recent list of autoimmune diseases associated with CD in the literature. This is of interest because research subjects evolve over time and in this light Medline acts as a biomedical research collective memory and an up-to-date view of clinical expertise. For example, based on MEDLINE co-occurrences before year 2000, *Autoimmune Hepatitis* would not be in the top 15 selection, but *Pemphigus* and *Pemphigoid, Bullous* would be (see Additional file 2). Combined with EHR mining give us prevalence estimate of comorbidities that were not suspected as being associated to CD at the time patients' diagnoses were recorded. Another advantage of this data-driven selection is to provide an automatable alternative to the usual elicitation step which classically determines relevant comorbidities by domain experts. This method allows to design more pragmatic studies, not relying on one or two experts' opinion.

Automatable and RECORD statement compliant workflow

The workflow of this study, i.e. comorbidities selection from the MeSH® co-occurrences file, mapping from MeSH® to other terminologies, and case identification through text and coded data mining, can be automated to estimate comorbidities burden in other EHR-based studies.

Our workflow is in line with RECORD statement [32], in particular reporting a complete list of codes and algorithms for each comorbidities.

Moreover, we demonstrated that manual review could be performed easily using text visualization tools integrated with the CDW, even for non-English language based EHR [29].

Phenotyping quality is sometimes considered as a limit of EHR reuse. In the proposed workflow, we reinforce phenotyping quality by manual phenotype extraction by two readers in reasonable time thanks to assisted extraction using a visualization tool, FASTVISU. While FASTVISU is based on regular expressions which lack of precision, our workflow could be improved using a natural language based tool.

Conclusion

We provide an automatable workflow fulfilling requirements from RECORD statement on observational routinely-collected health data aiming at identifying comorbidities burden for a specific disease using EHR. We applied this workflow to finely phenotype autoimmune comorbidities in a large CD population. We think that this flexible workflow will ease the extraction of relevant information from EHR.

Additional files

Additional file 1: List of ATC codes used. Lists of Anatomical Therapeutic Chemical Classification System (ATC) codes used for autoimmune thyroiditis (levothy*) and for diabetes mellitus, Type 1 (insulin). (DOCX 13 kb)

Additional file 2: Evolution of the numbers of co-occurrences in time. The 15 first ranked autoimmune diseases (in red) which would have been included based on the literature available at various time points. Numbers of co-occurrences until the specified year, ranks in prevalence estimates from this study, ranks in number of MeSH terms co-occurrence with term 'Celiac Disease' in MEDLINE at specified years. First version of the clinical vignette related on a new analgesic to control pain in mild trauma injuries with the four experimental factors tested. Description of first clinical vignette and list of response options. (DOCX 17 kb)

Acknowledgments
David Baudoin participated in the data management of the list of CD patients.
Dr. Eric Zapletal participated in the CDW data extractions.
Dr. Pierre Durieux revised critically the manuscript.

Funding
No external funding, apart from the support of the authors' institution, was available for this study.

Availability of data and materials
MeSH®, ATC and MRCOC file are available in the UMLS from the U.S. National Library of Medicine.
Data generated from these terminologies and dataset during this study are included in this published article and its Additional files, with exception of individual clinical data.

Authors' contributions
JBE conceived the study, reviewed the EHR, conducted the data analysis, participated in data interpretation and redacted the manuscript. BR participated in the study design, the terminologies extraction, data interpretation, drafting the work. GM participated in the study design and the data interpretation. SK participated in the study population inclusion, and the data interpretation. AB participated in the study design, the data interpretation, and drafting the work. CC participated in the data interpretation. ASJ conceived the study, reviewed the EHR, and participated in data interpretation and in the manuscript redaction. Every authors revised critically the manuscript, gave their final approval and agreed to be accountable for all aspects of the study.

Ethics approval and consent to participate
This study has been approved by the institutional review board for observational study (*Comité éthique de recherche des hôpitaux universitaires Paris Ouest*) under registration #00001072.
Patient consent for reuse of their EHR on an opt-out basis is declared at the *Commission Nationale de l'Informatique et des Libertés,* references #1743502 and #1695855v0.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Georges Pompidou European Hospital (HEGP), AP-HP, Paris, France. [2]INSERM UMRS 1138, Paris Descartes University, Paris, France. [3]Pôle Informatique Médicale et Santé Publique, Hôpital Européen Georges Pompidou, 20 rue Leblanc, 75015 Paris, France.

Received: 18 November 2016 Accepted: 12 September 2017
Published online: 29 September 2017

References

1. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: a 8-years follow-up experience. Int J Med Inform. 2017;102:21–8.
2. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc. 2014;21:221–30.
3. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. AMIA Annu Symp Proc. 2011;2011:274–83.
4. Benchimol EI, Guttmann A, Mack DR, Nguyen GC, Marshall JK, Gregor JC, et al. Validation of international algorithms to identify adults with inflammatory bowel disease in health administrative data from Ontario, Canada. J Clin Epidemiol. 2014;67:887–96.
5. Bertaud V, Lasbleiz J, Mougin F, Burgun A, Duvauferrier R. A unified representation of findings in clinical radiology using the UMLS and DICOM. Int J Med Inf. 2008;77:621–9.
6. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000;7:593–604.
7. Hahn U, Romacker M, Schulz S. MEDSYNDIKATE–a natural language system for the extraction of medical information from findings reports. Int J Med Inf. 2002;67:63–74.
8. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004;11:392–402.
9. Bakken S, Hyun S, Friedman C, Johnson SB. ISO reference terminology models for nursing: applicability for natural language processing of nursing narratives. Int J Med Inf. 2005;74:615–22.
10. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc. 2008;2008:404–8.
11. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc. 2011;2011:1564–72.
12. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 2016; 23:e20–7.
13. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23(6):ocv202.
14. Cosnes J, Cellier C, Viola S, Colombel J, Michaud L, Sarles J, et al. Incidence of autoimmune diseases in celiac disease: protective effect of the gluten-free diet. Clin Gastroenterol Hepatol. 2008;6:753–8.
15. Iqbal T, Zaidi MA, Wells GA, Karsh J. Celiac disease arthropathy and autoimmunity study. J Gastroenterol Hepatol. 2013;28:99–105.
16. Collin P, Salmi J, Hällström O, Reunala T, Pasternack A. Autoimmune thyroid disorders and coeliac disease. Eur J Endocrinol Eur Fed Endocr Soc. 1994; 130:137–40.
17. Diamanti A, Ferretti F, Guglielmi R, Panetta F, Colistro F, Cappa M, et al. Thyroid autoimmunity in children with coeliac disease: a prospective survey. Arch Dis Child. 2011;96:1038–41.
18. van der Pals M, Ivarsson A, Norström F, Högberg L, Svensson J, Carlsson A. Prevalence of thyroid autoimmunity in children with celiac disease compared to healthy 12-year olds. Autoimmune Dis. 2014;2014:417356.
19. Sategna-Guidetti C, Volta U, Ciacci C, Usai P, Carlino A, De Franceschi L, et al. Prevalence of thyroid disorders in untreated adult celiac disease patients and effect of gluten withdrawal: an Italian multicenter study. Am J Gastroenterol. 2001;96:751–7.
20. Counsell CE, Taha A, Ruddell WS. Coeliac disease and autoimmune thyroid disease. Gut. 1994;35:844–6.
21. Lubrano E, Ciacci C, Ames PR, Mazzacca G, Oriente P, Scarpa R. The arthritis of coeliac disease: prevalence and pattern in 200 adult patients. Br J Rheumatol. 1996;35:1314–8.
22. Volta U, Caio G, Stanghellini V, De Giorgio R. The changing clinical profile of celiac disease: a 15-year experience (1998-2012) in an Italian referral center. BMC Gastroenterol. 2014;14:194.
23. Størdal K, Bakken IJ, Surén P, Stene LC. Epidemiology of Coeliac Disease and Comorbidity in Norwegian Children: J. Pediatr Gastroenterol Nutr. 2013;57:467–71.
24. Bybrant MC, Örtqvist E, Lantz S, Grahnquist L. High prevalence of celiac disease in Swedish children and adolescents with type 1 diabetes and the relation to the Swedish epidemic of celiac disease: a cohort study. Scand J Gastroenterol. 2014;49:52–8.
25. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform. 2010;160:193–7.
26. Al-Hussaini A, Sulaiman N, Al-Zahrani M, Alenizi A, El Haj I. High prevalence of celiac disease among Saudi children with type 1 diabetes: a prospective cross-sectional study. BMC Gastroenterol. 2012;12:180.
27. Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. Brief Bioinform. 2016;17:33–42.
28. Abdelali B, Caruba T, Zapletal E, Sabatier B, Durieux P, Degoulet P. A Clinical Data Warehouse-Based Process for Refining Medication Orders Alerts. J Am Med Informat Assoc: JAMIA. 2012;19(5):782–85. doi:10.1136/amiajnl-2012-000850.
29. Escudié J-B, Jannot A-S, Zapletal E, Cohen S, Malamut G, Burgun A, et al. Reviewing 741 patients records in two hours with FASTVISU. AMIA Annu Symp Proc. 2015;2015:553–9.
30. Sperrin M, Thew S, Weatherall J, Dixon W, Buchan I. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. AMIA Annu Symp Proc. 2011;2011:1318–25.
31. Casez P, Labarère J, Sevestre M-A, Haddouche M, Courtois X, Mercier S, et al. ICD-10 hospital discharge diagnosis codes were sensitive for identifying pulmonary embolism but not deep vein thrombosis. J Clin Epidemiol. 2010; 63:790–7.
32. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med. 2015 [cited 2016 Oct 28];12 Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4595218/.
33. Hruby GW, Matsoukas K, Cimino JJ, Weng C. Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics. J Biomed Inform. 2016;60:376–84.
34. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. Health Aff Proj Hope. 2015;34:2174–80.