Atopic dermatitis or eczema? Consequences of ambiguity in disease name for biomedical literature mining

Clément Frainay, Yoann Pitarch, Sarah Filippi, Marina Evangelou, Adnan Custovic

Received: 9 March 2021 | Accepted: 30 June 2021

DOI: 10.1111/cea.13981

ORIGINAL ARTICLE

Clinical Allergy

WILEY

Atopic dermatitis or eczema? Consequences of ambiguity in disease name for biomedical literature mining

Clément Frainay[1,2] | Yoann Pitarch[3] | Sarah Filippi[4] | Marina Evangelou[1,4] | Adnan Custovic[5]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, London, UK

[2]Toxalim (Research Center in Food Toxicology), INRAE, ENVT, INP-PURPAN, UPS, Université de Toulouse, Toulouse, France

[3]UMR5505, IRIT, Université de Toulouse, Toulouse, France

[4]Department of Mathematics, Faculty of Natural Sciences, Imperial College London, London, UK

[5]National Heart and Lung Institute, Imperial College London, London, UK

Correspondence
Adnan Custovic, National Heart and Lung Institute, Imperial College London, London, UK.
Email: a.custovic@imperial.ac.uk

Funding information
INRAE; Medical Research Council

Abstract

Background: Biomedical research increasingly relies on computational approaches to extract relevant information from large corpora of publications.

Objective: To investigate the consequence of the ambiguity between the use of terms "Eczema" and "Atopic Dermatitis" (AD) from the Information Retrieval perspective, and its impact on meta-analyses, systematic reviews and text mining.

Methods: Articles were retrieved by querying the PubMed using terms 'eczema' (D003876) and "dermatitis, atopic" (D004485). We used machine learning to investigate the differences between the contexts in which each term is used. We used a decision tree approach and trained model to predict if an article would be indexed with eczema or AD tags. We used text-mining tools to extract biological entities associated with eczema and AD, and investigated the discrepancy regarding the retrieval of key findings according to the terminology used.

Results: Atopic dermatitis query yielded more articles related to veterinary science, biochemistry, cellular and molecular biology; the eczema query linked to public health, infectious disease and respiratory system. Medical Subject Headings terms associated with "AD" or "Eczema" differed, with an agreement between the top 40 lists of 52%. The presence of terms related to cellular mechanisms, especially allergies and inflammation, characterized AD literature. The metabolites mentioned more frequently than expected in articles with AD tag differed from those indexed with eczema. Fewer enriched genes were retrieved when using eczema compared to AD query.

Conclusions and Clinical Relevance: There is a considerable discrepancy when using text mining to extract bio-entities related to eczema or AD. Our results suggest that any systematic approach (particularly when looking for metabolites or genes related to the condition) should be performed using both terms jointly. We propose to use decision tree learning as a tool to spot and characterize ambiguity, and provide the source code for disambiguation at https://github.com/cfrainay/ResearchCodeBase.

KEYWORDS
atopic dermatitis, eczema, information retrieval, medical terminology, text mining

1 | INTRODUCTION

Investigations of skin conditions characterized by itchy rashes span several centuries, but are still subject to ambiguous terminology and definitions.[1] The study of a disorder that was initially termed "Eczema" has been marked through time by the proposition of a number of alternative denominations. As our understanding was growing, new terms emerged, resulting in the coexistence of several disease names and a growing ambiguity regarding the definitions. Despite numerous efforts to reach a consensus, two names currently coexist and are widely used: *Eczema* and *Atopic Dermatitis* (AD).[2] These terms are often used interchangeably and described as synonyms, as in the Online Mendelian Inheritance in Man (OMIM) database.[3] AD is frequently considered to be a more "specific" type of eczema, as in the Disease Ontology[4] (where eczema is considered as a synonym of dermatitis). This view is not supported by the World Allergy Organization,[5] but is used by the International Statistical Classification of Diseases and Related Health Problems (ICD) nomenclature. A further term *atopic eczema* was coined, but is used less frequently.[2,6] The term *AEDS*, for atopic eczema/dermatitis syndrome has also been proposed.[7] The 'atopic' qualifier originally represented a link with type I hypersensitivity,[8] but atopic sensitization itself was shown to be heterogeneous.[9,10] Recent studies have emphasized the need for a consensus terminology[2,6,11] and warned that ambiguity could jeopardize treatment reimbursement, patient education, as well as data mining. Similar pleas for terminology harmonization can be traced back to 1951.[12] Thus, decades of research with more than 25,000 articles in PubMed could have been affected by this terminology issue.

Information Retrieval (IR) is a field of research aiming to develop methods to obtain relevant information from large collections of information sources. It has gained considerable interest over the past decades of the exponential growth of information available on-line, which has led to search engines becoming key tools and one of the main entry points to scientific knowledge.[13] Scientists rely on online platforms such as PubMed and their dedicated search engine to cover the findings related to their field.[14] IR is of particular interest for meta-analyses and systematic reviews, where a comprehensive search for the relevant information is the first and critically important step. Disease names are among the most used terms for querying the PubMed database,[14,15] underlying the critical issue of eczema/AD ambiguity regarding IR. Beyond finding relevant documents, a considerable effort has been invested into automatically extracting information from the published literature, facilitating the extraction of list of genes, proteins or metabolites.[16–18] Moreover, other methods can extract the relationship between biological entities cited in texts, allowing automatic reconstruction of regulatory networks or protein–protein interactions to identify disease pathways.[19] These techniques fall under the classification of Text Mining techniques, which aim to process and extract information automatically from text documents. They usually rely on Natural Language Processing (NLP), and are commonly used in IR context. These applications rely on thesauri and ontology to resolve semantic ambiguity and grasp relationship between concepts. One of the resources commonly used is the Medical Subject Headings (MeSH) thesaurus,[20] which provides a controlled biomedical vocabulary, hierarchically structured and machine-readable, that describes the concepts in biomedical sciences. It is notably used for annotating and indexing articles in the MEDLINE database. The annotation is manually performed by experts reviewing the article content. The MeSH thesaurus constitutes a key component of the PubMed search engine: According to PubMed documentation, each untagged word in a PubMed query is first compared to a MeSH translation table to ensure the extraction of the relevant articles. The processing of user's query for Automatic Term Mapping (ATM) and query extension makes the MeSH critically important for document retrieval. In the MeSH thesaurus, AD and eczema are considered as two separated entities (D003876 and D004485), with no hierarchical relationship – that is, they are considered as two distinct concepts lying on the same level of the hierarchal structure, under the broader concept of 'Dermatitis'. Using one term or the other could consequently have a strong impact on the retrieved documents when querying PubMed.

Key Message

- Despite being used interchangeably, eczema and atopic dermatitis carry different findings and cover different topics in their respective corpora retrieved from PubMed.
- Any systematic analysis of eczema or atopic dermatitis literature, especially text mining approaches extracting associated genes and molecules should include both terms as input to account for the historical ambiguity.
- The feature extraction provided by our Decision Tree approach can help disambiguate AD/eczema-related articles, and support better query design for information retrieval.

The aim of this study was not to tackle which term (eczema or AD) is more appropriate, but to investigate the potential consequence of the ambiguity from the IR perspective, and its potential impact on meta-analyses, systematic reviews and text mining. Through a systematic characterization of the context of use of each term, using text mining techniques, we provide insights regarding the bias stemming from the choice of terminology.

2 | METHODS

2.1 | Finding the topics associated with a biomedical term

We first used Web of Science journal categories to cover the main research fields and related communities associated with the use of each term. Relevant articles related to eczema or AD from 1945 to

2017 were retrieved by querying the PubMed search engine using the terms *eczema* (D003876) and *dermatitis*, *atopic* (D004485) from the MeSH. According to the search engine documentation, all MeSH term that sits below those terms in the hierarchy were also included in the search. For *eczema*, this implies that the results include articles related to *dyshidrotic eczema*. However, other "*Eczematous skin diseases*" such as *contact eczema* or *seborrheic dermatitis* are not included as they are considered "sibling" terms in the MeSH thesaurus. We choose the corpus that pre-dates the recent 2017 recommendations on terminology,[1,2] as their endorsement and impact cannot be properly assessed yet at the time of writing this article. We analysed the indexing policy of the eczema/AD related literature in PubMed by performing a trend analysis similar to that used in the recent systematic review and meta-analysis.[2] We focused on the database lookup strategy set by the platform rather than the direct user interaction with it, although both are closely related and yield similar results.

Beyond query building, we used MeSH terms to describe other topics covered by eczema and AD articles. To characterize these articles and identify any differences, we extracted other MeSH terms associated with each of the two corpora. Smalheiser and Bonifield recently proposed a metric for quantifying the semantic relatedness of two MeSH terms through their tendency to co-occur in the same article's MEDLINE entry.[21] We used this metric and ranked associated MeSH terms according to their odd ratio and kept the top 40 associated MeSH terms (Table S1).

2.2 | Predicting indexing from abstract and title content

To grasp the differences between the two concepts, we trained a model to predict if an article would be indexed in PubMed with eczema or AD tags, using the content of the title and the abstract. We choose a decision tree approach to extract important terms that distinguish eczema from AD articles. The algorithm builds a comprehensive set of rules to classify data, from a learning set given as input. In our dataset, the instances are publications. The features used to describe them are the word content of the title and abstract. It is represented in the form of a high-dimensional binary vector where each position represents a word, and their value represents whether the word is present or not in the document. The class to predict was the PubMed annotation of the document, *eczema* or *AD* (documents matching both terms were excluded). The algorithm splits the learning set according to each feature and selects at each step the one that yields the best class separation. The process is repeated recursively until a fixed depth is reached or when the size of the remaining set is below a given threshold.

Identifiers and documents were retrieved using the PubMed REST API, allowing programmatic access to the database content. The abstract and title were pre-processed in order to remove stop words and harmonize vocabulary as lower case lemma, using nltk's WordNet lemmatizer.[22,23] The lemmatizer allows to collapse different inflectional forms, for example, *mouse* and *mice*, as one single feature. To avoid obvious classification rules, the terms *dermatitis*, *atopic* and *eczema* were filtered. We used the CART (Classification And Regression Trees) implementation of the Scikit-learn python library[24] and Gini impurity as the splitting criterion. As there were more AD than eczema articles, the learning set was balanced using random sampling of the main class to avoid bias. We used a maximum depth of 6 and a minimum sample size of 1% to avoid over-fitting. We performed cross-validation to assess the quality of the model, keeping 20% of the dataset off during the learning phase.

2.3 | Extracting biological entities associated with a biomedical term

Bio-entities associated with eczema and AD were extracted using text-mining software which scans a large corpus of documents and performs Named Entity Recognition (NER) to detect mention of biological entities or use annotations from curated database. This process is followed by a statistical analysis to select the biological entities that best characterize the corpus. We used Polysearch[25] and Gene Set to Diseases (GS2D)[26] to find enriched protein-coding genes significantly associated with a set of articles indexed with a particular MeSH term (AD or eczema). Enriched compounds were retrieved using Polysearch[25] and Metab2MeSH.[27] We also used Alkemio[28] and Génie,[29] which use the MeSH-indexed documents to build a model characterizing the topic, in order to extend the considered corpora beyond documents indexed under the given MeSH terms. For each tool, we used default cut-offs proposed by the developers. Details of each tool and setting can be found in Appendix S1.

3 | RESULTS

3.1 | Trend analysis of the eczema and Atopic Dermatitis terms use over time

The terms "eczema" and "AD" were rarely used jointly for annotating articles (only 4.7% of the total articles in PubMed published between 1945 and the end of 2017), leading to the retrieval of different documents if only one term was used in the query. *AD* is more recent term than *eczema*, its first appearance dates from 1933.[30] While rarely used until the mid-1960s, AD has gradually overtaken eczema as a preferred indexing term among all articles in PubMed, particularly over the last two decades (Figure 1). Among 15,287 related articles in PubMed published between 1995 and 2015, 12,262 (80%) are indexed under AD. The term *eczema* has regained some popularity since the 1990s, with a growing number of articles indexed under this term. Figure 1 suggests that a stable balance has been reached during the last decade, with a ratio of 4:1 in favour of AD.

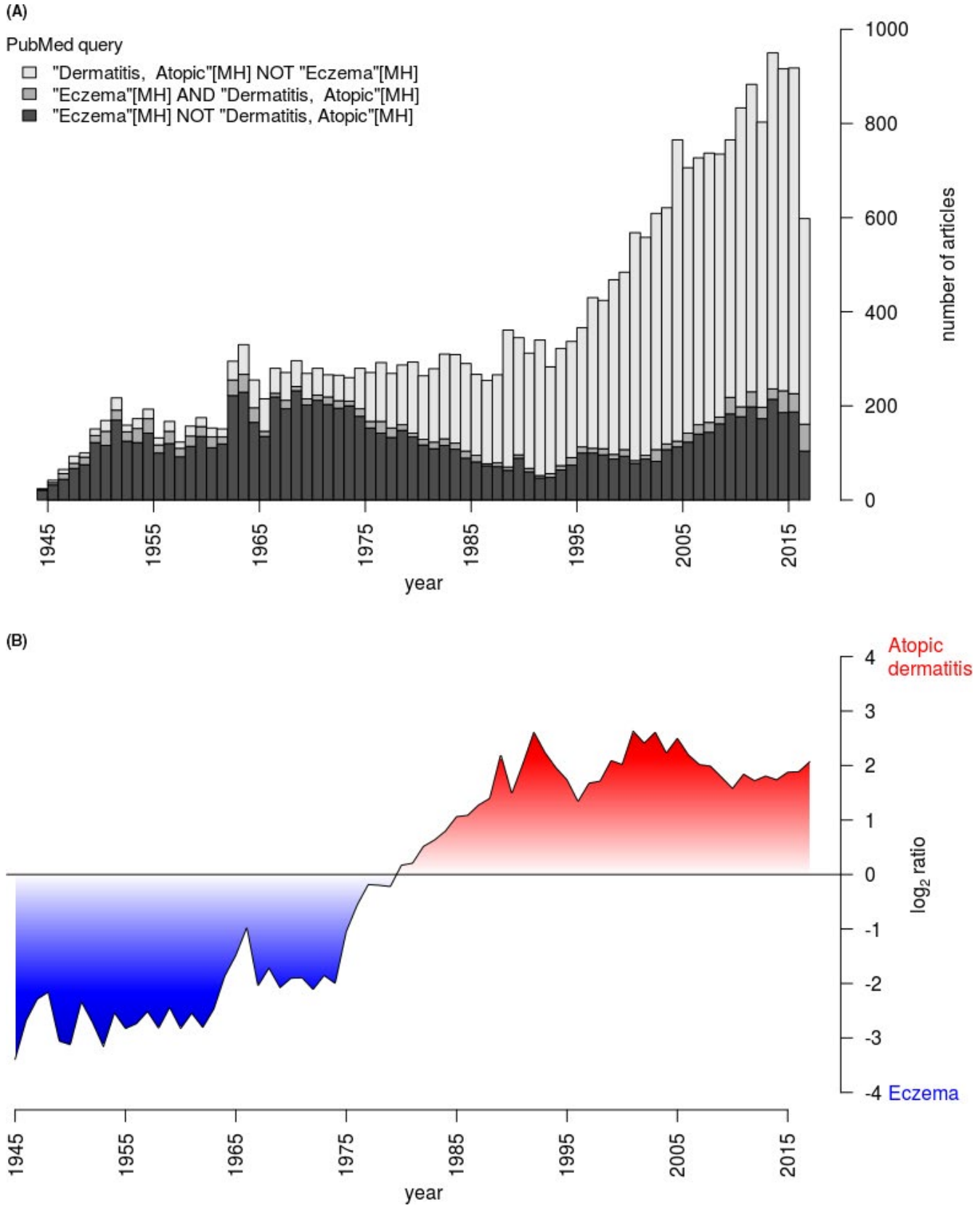

FIGURE 1 Atopic dermatitis and eczema indexing trends in PubMed from 1945 to 2017.(A) Number of articles by years retrieved from PubMed Search Engine using eczema or atopic dermatitis MeSH terms as query terms. Stacked histograms coloured according to query used. (B) Log² ratio of articles in PubMed retrieved from atopic dermatitis MeSH term over articles retrieved from eczema MeSH term

3.2 | Distribution of Eczema and Atopic Dermatitis articles among scientific fields

Atopic dermatitis query yielded more articles related to veterinary science, biochemistry, cellular and molecular biology; the eczema query linked to a larger proportion of public health, infectious disease and respiratory system articles (Figure 2).

3.3 | Analysis of the topics associated with Eczema and Atopic Dermatitis articles

We extracted the list of MeSH terms used jointly with *AD* or *eczema* more frequently than would be expected by chance. As expected, *eczema* and *AD* are related given such criterion, appearing in each other list of related terms. However, MeSH terms frequently associated with "Atopic, Dermatitis" or "Eczema" differed, with an agreement between the unordered top 40 lists of 52% (Figure 3). Food hypersensitivity and IgE are strongly related to AD, while in contrast, eczema shared many connections with other types of dermatitis, especially "neurodermatitis".

3.4 | Disaggregation of Eczema and Atopic Dermatitis articles using machine learning

We applied a machine-learning algorithm (decision tree learning) to create a model to distinguish *AD* from *eczema* articles and extract important features that could help narrow the definition and context of use of both terms (Figure 4). The presence of the word "*cell*" in an abstract was enough to extract a substantial part of our training set from the literature on the topic (10.7%) with an over-representation of AD articles (83.4%). This proportion was increased if the immunity-related word such as "*inflammatory*", "*cytokine*" or "*IgE*" are present in the abstract or the title. However, in the absence of such terms, an article can still be assigned to the AD class. The presence of the word "*dog*" assigned AD label with a decent confidence, but to a small portion of the literature. The presence of the word "*child*" also tended to be frequent in non-immunology related AD articles, leading to the most "uncertain" leaf of the decision tree (which represented 9.3% of our samples for which the class attribution was close to a random guess). In our classification model, the presence of specific words positively contributed to classify a document as AD indexed. In contrast, the assignation to the eczema class was mainly

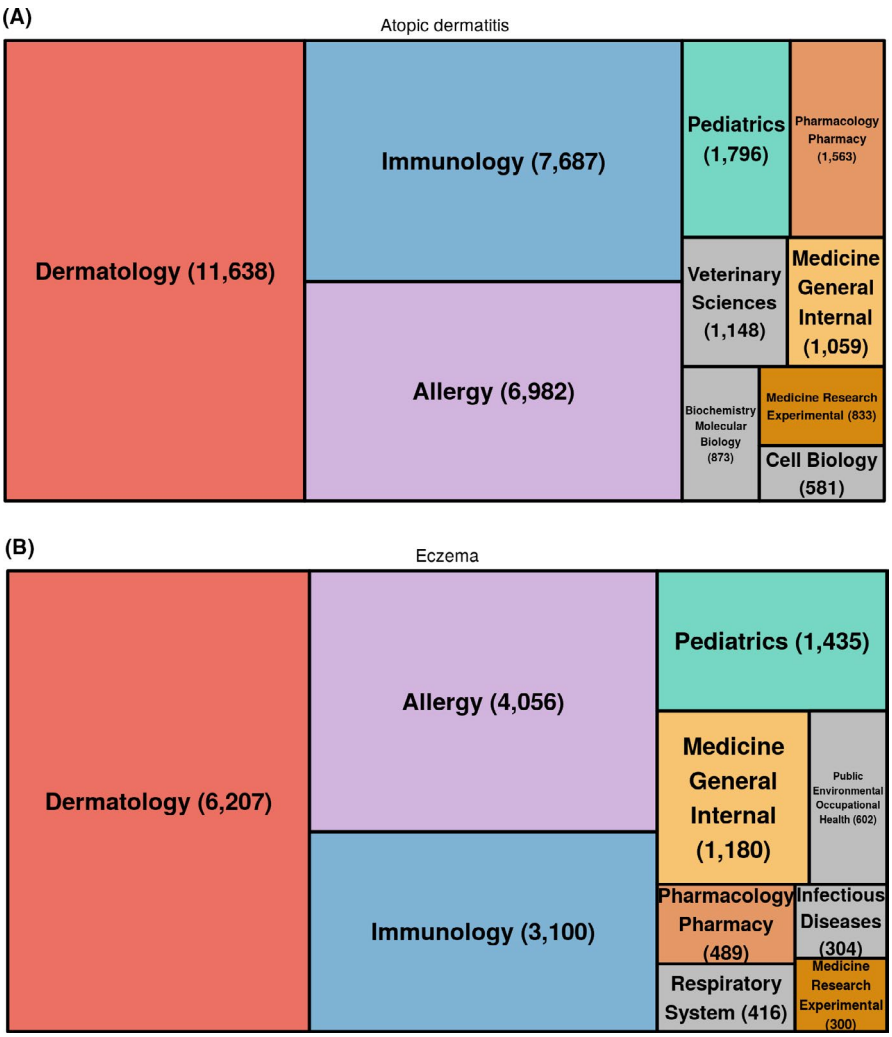

FIGURE 2 Treemap of repartition of articles retrieved from eczema and atopic dermatitis queries in Web of Science's categories (top 10th category for each query). Documents from 1956 to 2017. Grey colour represent category not shared between the two top 10th. Tile area proportional to the number of articles. (A) Category proportion for atopic dermatitis query (B) Category proportion for eczema query

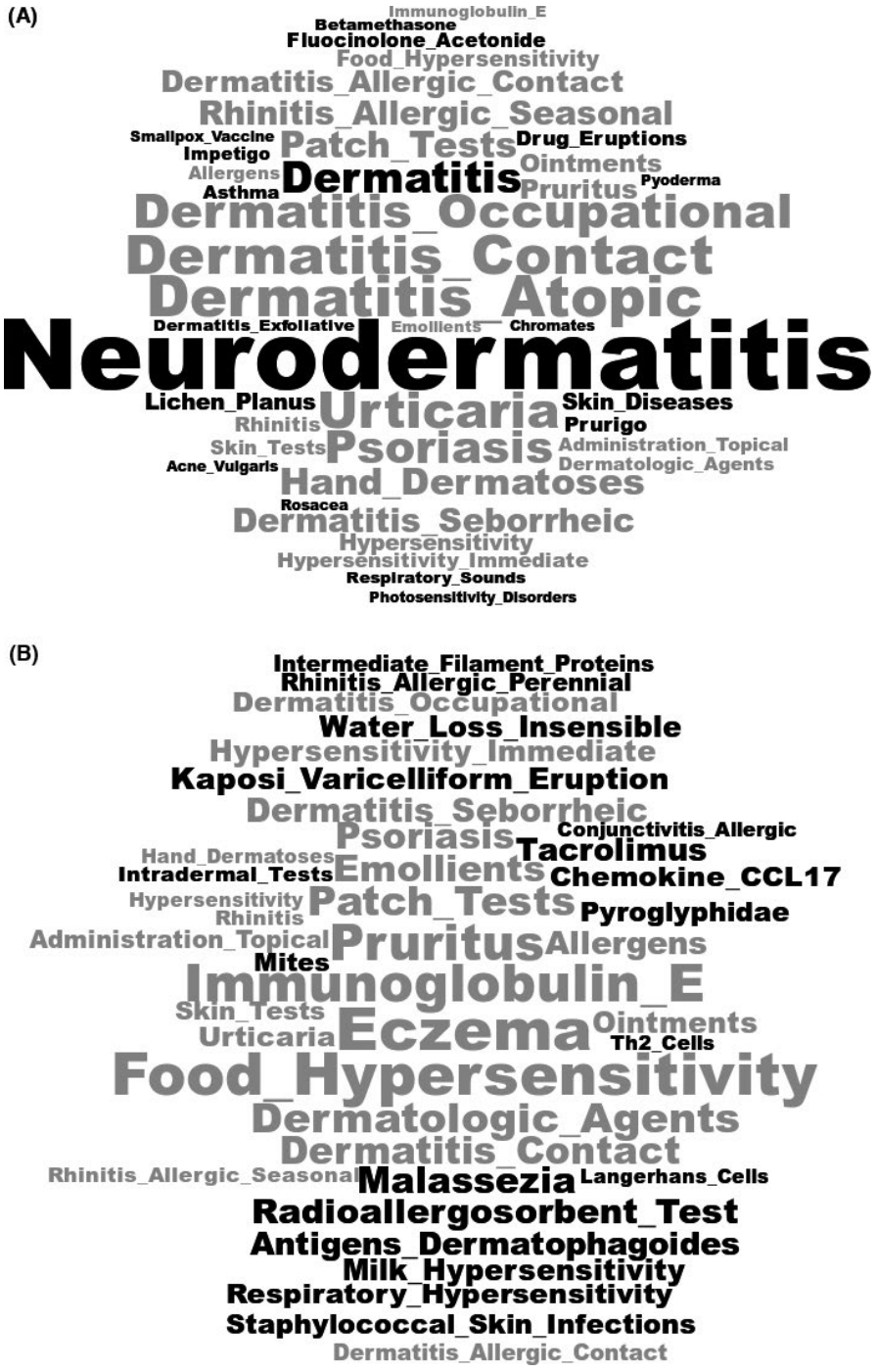

FIGURE 3 MeSH terms clouds representing the top 40th terms associated with eczema and Dermatitis, atopic, according to Smalheiser and Bonifield 'article' similarity for semantic relatedness. Term size proportional to odds ratio. Terms depicted in black represent terms not shared between the two lists. (A) MeSH cloud associated with eczema terms. (B) MeSH cloud associated with Dermatitis, atopic terms

driven by the absence of those terms, with only two words whose presence would be characteristic of the class. One of the two terms whose presence supported an assignation to the eczema class was the word "*hand*".

The precision for AD class was 0.8 and the recall of 0.66 when confronted to an unknown test set of articles. The precision for eczema class was 0.53 for a recall of 0.70.

The decision tree learning can be used as a tool to characterize ambiguity between the terms AD and eczema and support query refinement. We provide the source code for disambiguation at https://github.com/cfrainay/ResearchCodeBase.

3.5 | Identification of genetic associates and pathways according to the terminology used

Figure 5 shows that the metabolites mentioned more frequently than expected in MEDLINE articles related to the term *AD* differ from those found in articles related to the term *eczema*. On average, only 41.6% of the overall retrieved compounds were shared between the two queries. The coverage of retrieved entities was more skewed for genes, where less enriched genes were retrieved using *eczema* query, comparing when querying *AD*. On average the genes retrieved from *AD* cover 91.2% of the total number of genes

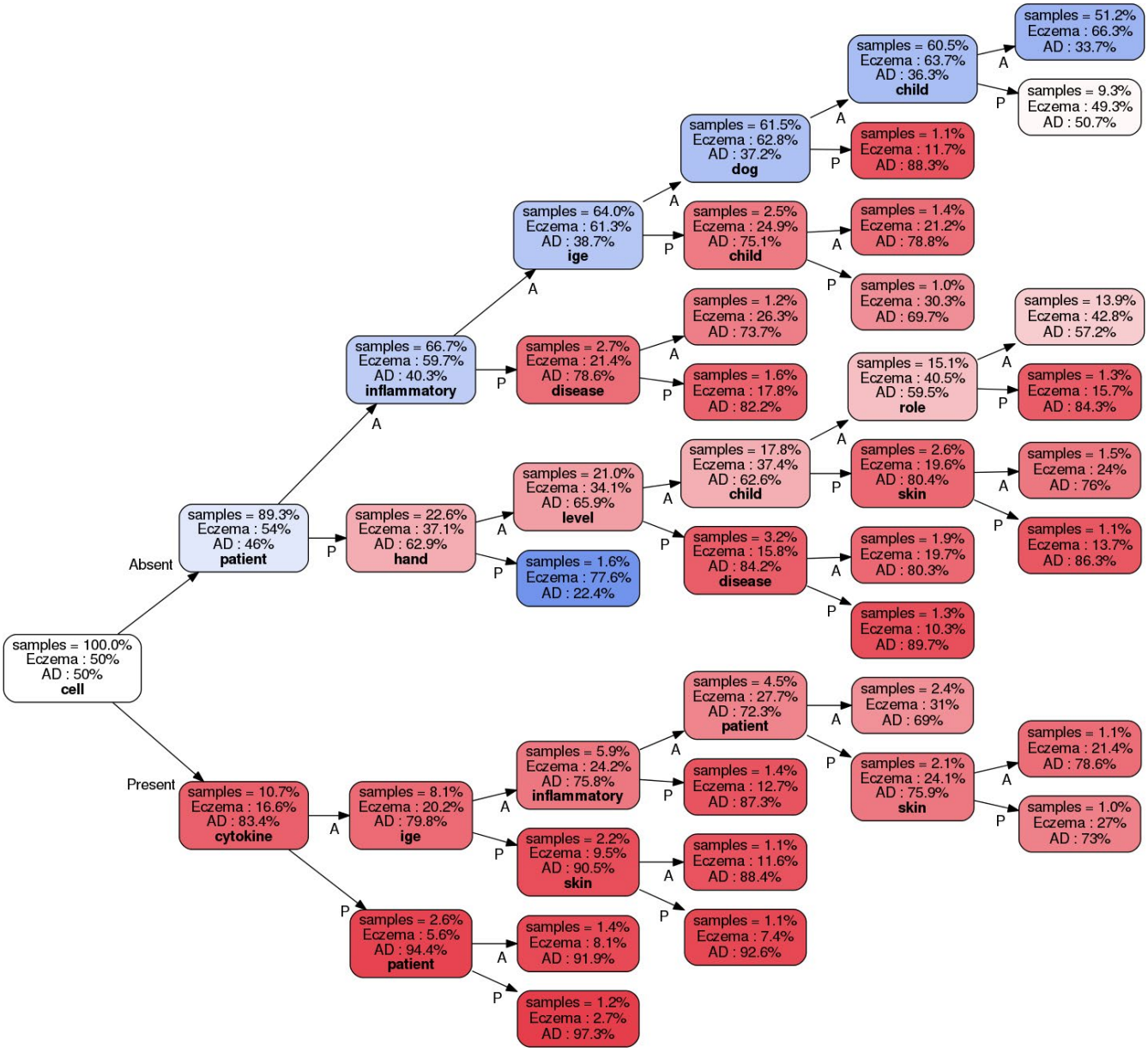

FIGURE 4 Decision tree for disaggregation of PubMed articles indexed with eczema or dermatitis, atopic MeSH terms. Node color represents assigned class, blue for eczema, red for atopic dermatitis. Shades intensity represents level of impurity (Gini), a measure of the quality of the split. Closer to white indicate a high level of impurity

retrieved from *eczema* or *AD*. GS2D retrieve only two genes from the *eczema* query, also retrieved from *AD* query. However, despite the paucity of genes associated with *eczema*, 17 genes retrieved from *eczema* search were not retrieved from *AD* search using Polysearch.

4 | DISCUSSION

Our findings suggest that the terms *eczema* and *atopic dermatitis* have been used in different contexts. By analysing the whole PubMed corpus from 1945 to 2017, we have shown that different names are bonded to different findings, which could impair systematic and automatic analysis of the literature. The literature associated with different aspects of the condition tends to have a preferred term, leading to bias when performing IR and extraction tasks. This may result in inconsistent findings when querying MeSH-indexed database such as PubMed, as shown using different automatic information extraction tools for genes and metabolites. While previous works supporting consensus denomination has warned about the consequence of such ambiguity on data mining,[2,11] this is, to our knowledge, the first systematic assessment of such consequences. Our results suggest that any systematic approach (particularly when looking for metabolites or genes related to the condition) should be performed using both terms jointly. Our model for distinguishing articles retrieved from *eczema* and from *AD* queries provides a model for refining a PubMed query.

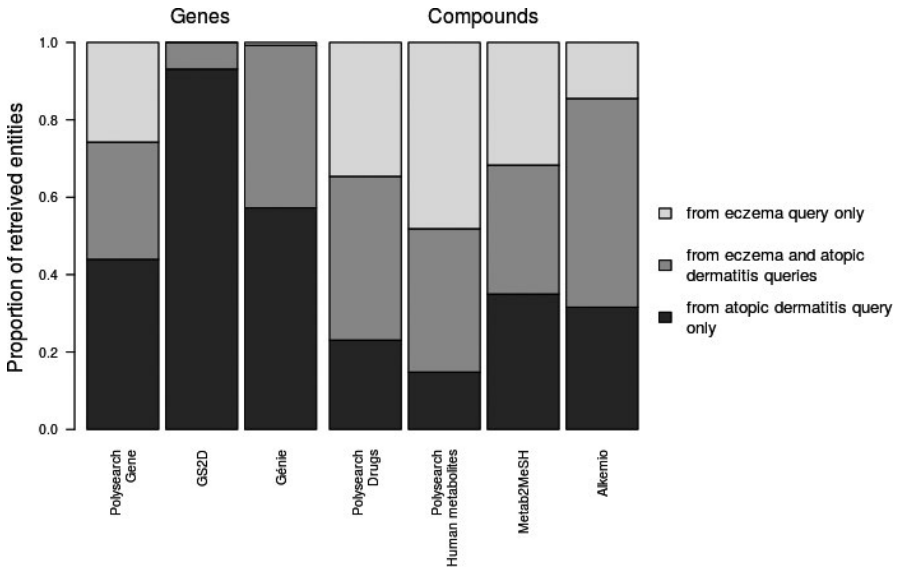

FIGURE 5 Results consistency of text-mining programs for relevant bio-entity retrieval when using atopic dermatitis or eczema as query. The central part of the stacked barplot represent the agreement, i.e. the proportion of bio-entities retrieved regardless of which of the two terms has been chosen as query

Kantor et al have shown that the prominence of each term differs between languages,[2] and that "*eczema*" was more used in French and German publications, whilst publications written in English used AD more often. Such results are hard to interpret given the fact that many publications in English are written by non-native speakers. However, this coincides with the history of the disease,[31] underlying the existence of American and European dermatology communities with different views regarding the nomenclature. Kantor et al[2] also showed that the distribution of term use varies between journals from different fields, namely dermatology, allergy, paediatrics and medicine. Our findings support these findings. We expanded the analysis by looking at the repartition of journals into Web of Science categories, which suggested an association between the use of AD and veterinary science, biochemistry, cellular and molecular biology. We assessed those differences at the topic level using articles annotations and text content, through a machine-learning classification approach. The model performance was fair for predicting AD indexing and supported the notion that the literature associated with each term is not homogeneous.

The classification of eczema articles lacked specificity, meaning that finding terms that characterize eczema and distinguish it from AD is difficult. In contrast, the classification of AD articles was specific but lacked sensitivity. It is therefore possible to find a subset of the AD literature, with a characteristic vocabulary not used in eczema articles. The decision tree and the tag clouds suggest that the presence of terms related to cellular mechanisms, especially allergies and inflammation, tends to characterize AD literature. The word "cell" seems to be an important criterion for distinguishing the two corpora, suggesting methodological preferences associated with each term. Those findings are consistent with the estimated pre-eminence of AD articles in cell and molecular science journals. One term whose presence can support an assignment to the eczema class was the word "hand". Hand eczema, or dyshidrotic eczema, refers to a more specific condition usually not coined as AD. It has its own MeSH entry, as a child node of eczema. In the absence of words related to immune system, the words "dog" tends to classify the document as AD-indexed. This is consistent with the topic analysis suggesting that AD is more used in veterinary science journals. The word "child" leads to a very uncertain leaf of the decision tree. This ambiguity could be related to the fact that infantile eczema is often referenced as AD. It is actually one of the synonyms listed in AD MeSH entry, but not in the eczema entry.

This heterogeneity of contexts associated with each term suggests that their selection for querying the PubMed database will result in articles that cover different topics and research focus. Consequently, the findings retrieved might differ according to the term chosen in the query, which would impact systematic literature analysis. This is supported by the results of text mining-based entity-extraction software, which shows limited agreement between the genes and compounds retrieved from AD and eczema queries.

Our results are an example of the implications of disease name ambiguity on text mining approaches, and emphasize the need to characterize, in terms of topics and content, the literature associated with each term and detect when two 'synonymous' disease names do not carry the same information. Although more sophisticated learning algorithms could be used to improve the prediction model accuracy, for example, gradient boosting decision trees,[32] we deliberately chose decision trees to favour the model interpretability. Decision tree learning has the advantage of offering an understandable output allowing one to clearly identify what drives the prediction of a class for a given instance. However, Decision Trees are known to be prone to over-fitting, and thus lack of robustness to small variations in the training set, especially regarding the deeper nodes with small sample size.

Although they can share some methodological aspects, our approach must be distinguished from the NLP task of Word Sens Disambiguation (WSD). WSD already attracted much attention in biomedical applications.[33,34] It aims at resolving other kinds of ambiguities, namely lexical ambiguity due to polysemy or homonymy, that is, for a word with different meanings, finding the right

one according to the context. An example would be to define if the word "capsule" refers to an anatomical cavity, a pharmaceutical product, a bacterial membrane, or a plant structure; or if the word "cat" refers to a species, the Computerized Axial Tomography or the Chloramphenicol acetyltransferase gene. Each of those concepts has a clearly defined meaning. On the other hand, the eczema/AD ambiguity that we focused on is related to vagueness of definitions rather than lexical ambiguity, and stems from the overlap of the meaning of the two words.

The issue regarding the retrieval of relevant documents for eczema or AD research goes beyond the definition of a consensus terminology. If the community follows the recommendation of avoiding the name *eczema* over AD in further research, previous findings coined with the term *eczema*, relevant for deciphering AD, might be overlooked by text mining approach and search engines. Aiming at doing genuinely cumulative science, findings predating the emergence of a consensus definition need to be taken into account by IR techniques.

Our results should raise awareness of the potential bias imputed to the term used when relying on text-mining approach and exemplify the importance of setting proper time frame and terms when querying publication database. We propose that the feature extraction provided by our decision tree approach can provide such terms to disambiguate AD/eczema-related queries, and that this approach can be applied to decipher the complex relationship between other biomedical closely related concepts, help build accurate query for secondary science and support prompt reaction to settle consensus denominations.

AUTHOR CONTRIBUTIONS

AC, SF, ME and CF conceived goals and aim of the presented study. CF carried out corpora analysis. YP and CF performed code implementation and computation for the disambiguation. YP implemented data collection from PubMed. All authors took part in the design of the methodology. All authors discussed the results and contributed to the final manuscript.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Clément Frainay https://orcid.org/0000-0003-4313-2786
Yoann Pitarch https://orcid.org/0000-0002-1508-5436
Sarah Filippi https://orcid.org/0000-0001-8652-358X
Marina Evangelou https://orcid.org/0000-0003-0789-8944
Adnan Custovic https://orcid.org/0000-0001-5218-7071

REFERENCES

1. Bieber T. How to define atopic dermatitis? *Dermatol Clin*. 2017;35:275-281. https://doi.org/10.1016/j.det.2017.02.001
2. Kantor R, Thyssen JP, Paller AS, et al. Atopic dermatitis, atopic eczema, or eczema? A systematic review, meta-analysis, and recommendation for uniform use of 'atopic dermatitis'. *Allergy*. 2016;71:1480-1485. https://doi.org/10.1111/all.12982
3. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33:D514-D517. https://doi.org/10.1093/nar/gki033
4. Schriml LM, Arze C, Nadendla S, et al. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40:940-946. https://doi.org/10.1093/nar/gkr972
5. Johansson SGO, Bieber T, Dahl R, et al. Revised nomenclature for allergy for global use: Report of the Nomenclature Review Committee of the World Allergy Organization, October 2003. *J Allergy Clin Immunol*. 2004;113:832-836. https://doi.org/10.1016/j.jaci.2003.12.591
6. Silverberg JI, Thyssen JP, Paller AS, et al. What's in a name?: Atopic dermatitis or atopic eczema, but not eczema alone. *Allergy*. 2017;72:2026-2030. https://doi.org/10.1111/all.13225
7. Johansson SGO, Bousquet J, Dreborg S, et al. A revised nomenclature for allergy An EAACI position statement from the EAACI nomenclature task force. *Allergy*. 2001;56:813-824. https://doi.org/10.1111/j.1398-9995.2001.00002.x-i1
8. Pepys J. Natural history of "atopy". *J Allergy Clin Immunol*. 1986;78(5):959-961.
9. Simpson A, Tan VY, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med*. 2010;181(11):1200-1206. https://doi.org/10.1164/rccm.200907-1101OC. Epub 2010 Feb 18 PMID: 20167852.
10. Custovic A, Custovic D, Kljaic Bukvic B, Fontanella S, Haider S. Atopic phenotypes and their implication in the atopic march. *Expert Rev Clin Immunol*. 2020;16(9):873-881. https://doi.org/10.1080/1744666X.2020.1816825. Epub 2020 Sep 16 PMID: 3285695911.
11. Bieber T. Why we need a harmonized name for atopic dermatitis / atopic eczema / eczema!. *Allergy*. 2016;71:1379-1380. https://doi.org/10.1111/all.12984
12. Linn LW. The eczema-dermatitis nomenclature problem. *Aust J Dermatol*. 1951;1:127-134. https://doi.org/10.1111/j.1440-0960.1951.tb01415.x
13. Khare R, Leaman R, Lu Z. Accessing biomedical literature in the current information landscape. *Methods Mol Biol*. 2014;1159:11-31. https://doi.org/10.1007/978-1-4939-0709-0_2
14. Dogan RI, Murray GC, Névéol A, et al. Understanding PubMed user search behavior through log analysis. *Database (Oxford)*. 2009;2009:bap018. https://doi.org/10.1093/database/bap018
15. Erskovic JORH, Anaka LENYT, Ersh WIH, et al. A Day in the Life of PubMed: Analysis of a Typical Day's Query Log. *J. Am. Med. Informatics Assoc*. 2007;14(212–220): https://doi.org/10.1197/jamia.M2191
16. Gonzalez GH, Tahsin T, Goodale BC, et al. Recent advances and emerging applications in text and data mining for biomedical discovery. *Brief Bioinform*. 2016;17:33-42. https://doi.org/10.1093/bib/bbv087
17. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. *Genome Biol*. 2005;6:224. https://doi.org/10.1186/gb-2005-6-7-224
18. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*. 2006;7:119-129. https://doi.org/10.1038/nrg1768
19. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*. 2004;5:1-13. https://doi.org/10.1186/1471-2105-5-147
20. Nelson SJ. Medical terminologies that work: The example of MeSH. I-SPAN 2009–10th Int. Symp Pervasive Syst Algorithms Networks. 2009;380–384: https://doi.org/10.1109/I-SPAN.2009.84
21. Smalheiser NR, Bonifield G. Two similarity metrics for medical subject headings (MeSH): an aid to biomedical text mining and author

name disambiguation. *J Biomed Discov Collab*. 2016;7:1-14. https://doi.org/10.5210/disco.v7i0.6654
22. Bird S, Loper E. NLTK: The Natural Language Toolkit. Proc. ACL 2004 Interact. poster Demonstr. Sess. 2004; 31. https://doi.org/10.3115/1219044.1219075
23. Miller GA. WordNet: a lexical database for English. *Commun ACM*. 1995;38:39-41. https://doi.org/10.1145/219717.219748
24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
25. Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic Acids Res*. 2015;43:535-542. https://doi.org/10.1093/nar/gkv383
26. Andrade-navarro MA, Fontaine JF. Gene Set to Diseases (GS2D): disease enrichment analysis on human gene sets with literature data. *Genomics Comput Biol*. 2016;2:1-7. https://doi.org/10.18547/gcb.2016.vol2.iss1.e33
27. Sartor MA, Ade A, Wright Z, et al. Metab2MeSH: annotating compounds with medical subject headings. *Bioinformatics*. 2012;28:1408-1410. https://doi.org/10.1093/bioinformatics/bts156
28. Gijón-Correas JA, Andrade-navarro MA, Fontaine J-F. Alkemio: association of chemicals with biomedical topics by text and data mining. *Nucleic Acids Res*. 2014;42:422-429. https://doi.org/10.1093/nar/gku432
29. Fontaine J-F, Priller F, Barbosa-silva A, et al. Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res*. 2011;39:455-461. https://doi.org/10.1093/nar/gkr246
30. Wise F, Sulzberger MB. *Year Book of Dermatology and Syphilology*. Year B. Dermatology Syphilol. Chicago Year B. Publ. 1933;38–39.
31. Taïeb A, Wallach D, Tilles G. The History of Atopic Eczema / Dermatitis. *Handbook of atopic eczema*. Berlin, Heidelberg: Springer; 2006;10–20. https://doi.org/10.1007/3-540-29856-8_2
32. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146-3154.
33. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Informatics Assoc*. 2002;9:621-636. https://doi.org/10.1197/jamia.M1101
34. Jimeno-yepes AJ, Aronson AR. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics*. 2010;11:1-12. https://doi.org/10.1186/1471-2105-11-569

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Frainay C, Pitarch Y, Filippi S, Evangelou M, Custovic A. Atopic dermatitis or eczema? Consequences of ambiguity in disease name for biomedical literature mining. *Clin Exp Allergy*. 2021;51:1185–1194. https://doi.org/10.1111/cea.13981