Check for updates

Accepted Article

DR CLÉMENT FRAINAY (Orcid ID : 0000-0003-4313-2786)

PROFESSOR ADNAN CUSTOVIC (Orcid ID : 0000-0001-5218-7071)

Article type : Original Article

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

Clément Frainay PhD[1,2], Yoann Pitarch PhD[3], Sarah Filippi PhD[4], Marina Evangelou PhD[1,4] and Adnan Custovic MD PhD FMedSci[5]

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

[2] INRAE, UMR1331, Toxalim, Research Centre in Food Toxicology, Toulouse, France

[3] Université de Toulouse, UMR5505, IRIT, 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

Corresponding author:

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/cea.13981

This article is protected by copyright. All rights reserved

Professor Adnan Custovic, National Heart and Lung Institute, Imperial College London, London, UK

Tel: +44 20 7594 3274; fax: +44 20 7594 3984;

Email: a.custovic@imperial.ac.uk

Funding: CF was supported by INRA (Institut National de la Recherche Agronomique) CJS Program fellowship

AC contribution funded in part by MRC Grant MR/S025340/1

Disclosure of potential conflict of interest: The authors declare that they have no conflict of interests.

Accepted Article

This article is protected by copyright. All rights reserved

Accepted Article

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: AD 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 (MeSH) 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: Eczema; Atopic dermatitis; Text Mining; Information Retrieval; Medical terminology

Key messages :

This article is protected by copyright. All rights reserved

Accepted Article

- 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.

Author contribution statement:

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.

Abbreviations used

AD:	Atopic Dermatitis
OMIM	Online Mendelian Inheritance in Man
ICD:	International statistical classification of diseases and related health problems
IR:	Information Retrieval
AEDS:	Atopic eczema/dermatitis syndrome
NLP:	Natural Language Processing
MeSH:	Medical Subject Headings
CART:	Classification And Regression Trees
NER:	Named Entity Recognition
GS2D:	Gene Set to Diseases

This article is protected by copyright. All rights reserved

Accepted Article

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 on-line 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

This article is protected by copyright. All rights reserved

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 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 – i.e., 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.

The aim of this study is 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.

This article is protected by copyright. All rights reserved

Accepted Article

METHODS

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 time of writing. We analyzed 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).

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,

This article is protected by copyright. All rights reserved

from a learning set given as input. In our dataset, the instances are publications. The features used to describe them are the words 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, e.g. *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.

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

This article is protected by copyright. All rights reserved

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 the supplementary appendix.

Accepted Article

This article is protected by copyright. All rights reserved

Accepted Article

RESULTS

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 '60s, 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 '90s, 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 favor of AD.

Distribution of *Eczema* and *Atopic Dermatitis* articles among scientific fields

AD 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).

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".

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

This article is protected by copyright. All rights reserved

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 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; our model can be used for refining a PubMed query, and we provide the source code for disambiguation at https://github.com/cfrainay/ResearchCodeBase .

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 retrieved from *eczema* or *AD*. GS2D retrieve only 2 genes from the *eczema* query, also retrieved from *AD* query. However, despite the paucity

This article is protected by copyright. All rights reserved

of genes associated with *eczema*, 17 genes retrieved from *eczema* search were not retrieved from *AD* search using Polysearch.

DISCUSSION

Our findings suggest that the terms *eczema* and *atopic dermatitis* have been used in different contexts. By analyzing 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 information retrieval 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.

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.* also showed that the distribution of term use varies between journals from different fields, namely dermatology, allergy, paediatrics and medicine[2]. 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

This article is protected by copyright. All rights reserved

performance was fair for predicting AD indexing and supported the notion that the literature associated with each term isn't 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 preeminence 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 don't

This article is protected by copyright. All rights reserved

carry the same information. Although more sophisticated learning algorithms could be used to improve the prediction model accuracy, e.g. gradient boosting decision trees[32], we deliberately chose decision trees to favor 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, i.e. 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

This article is protected by copyright. All rights reserved

biomedical closely related concepts, help build accurate query for secondary science and support prompt reaction to settle consensus denominations.

Accepted Article

This article is protected by copyright. All rights reserved

Accepted Article

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—7, 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, Svensén M, Bishop CM, Heckerman DE, Buchan I, Custovic A. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. Am J Respir Crit Care Med. 2010 Jun 1;181(11):1200-6. doi: 10.1164/rccm.200907-1101OC. Epub 2010 Feb 18. PMID: 20167852

This article is protected by copyright. All rights reserved

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 Sep;16(9):873-881. doi: 10.1080/1744666X.2020.1816825. Epub 2020 Sep 16. PMID: 3285695911. 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 2009; 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

This article is protected by copyright. All rights reserved

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; 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

This article is protected by copyright. All rights reserved

30. Wise F, Sulzberger MB. Year book of dermatology and syphilology. Year B. Dermatology Syphilol. Chicago Year B. Publ. 1933; 38–39, https://doi.org/10.1001/archderm.1934.01460130180026

31. Taïeb A, Wallach D, Tilles G. The History of Atopic Eczema / Dermatitis. Handb. atopic eczema 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. 30 2017; 3146–3154

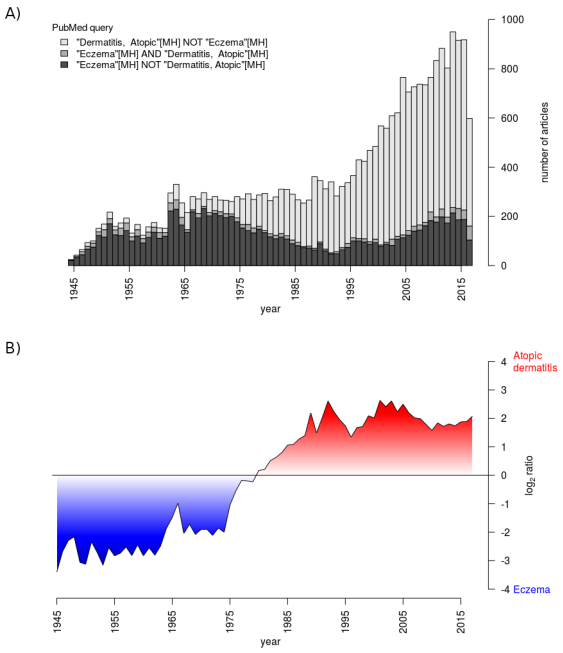
33. Liu H, Johnson SB, Friedman C. Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS. J. Am. Med. Informatics Assoc. 2002; 9:621–636, https://doi.org/10.1197/jamia.M1101

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

Accepted Article

This article is protected by copyright. All rights reserved

Accepted Article

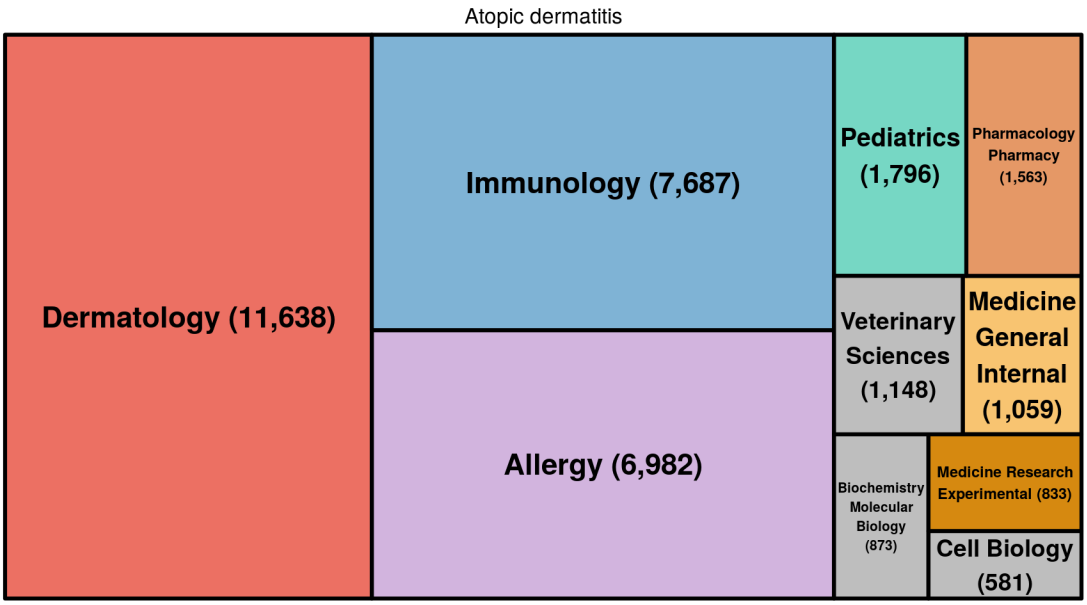


cea_13981_f1.tif

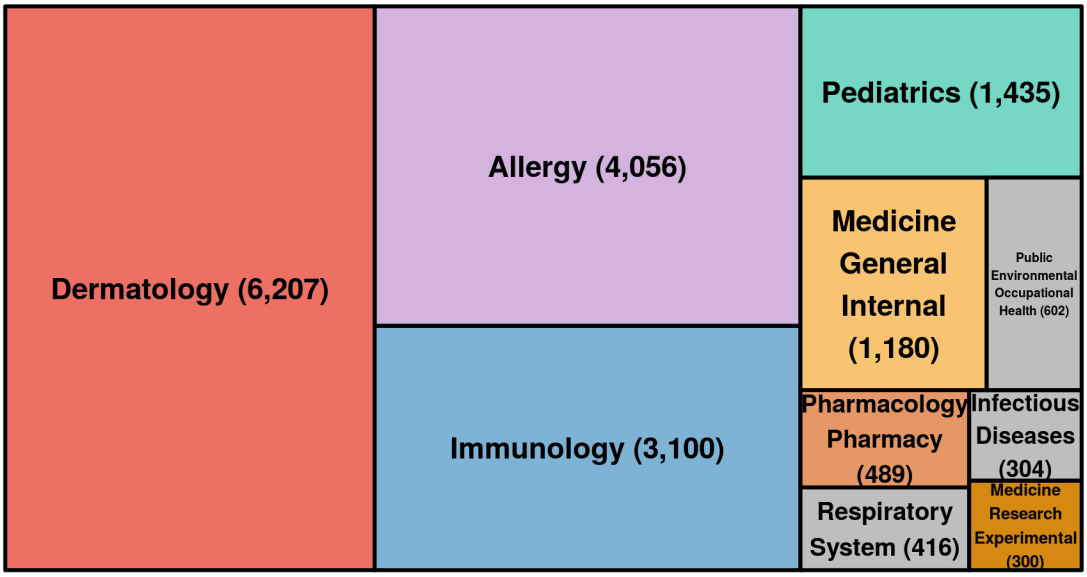
This article is protected by copyright. All rights reserved

Accepted Article

A)



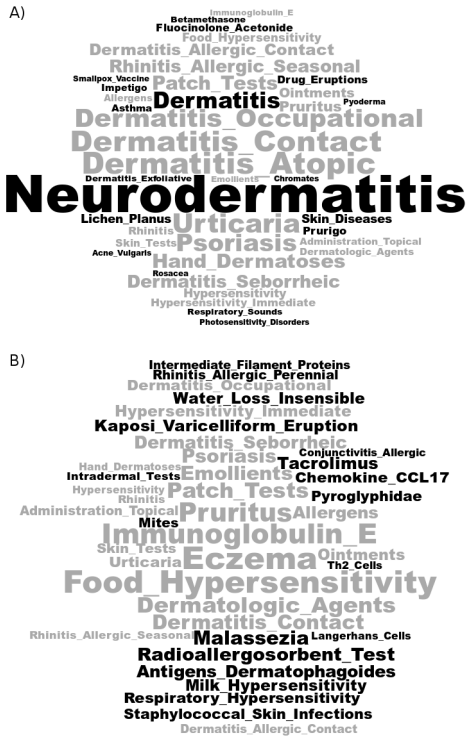
B)



cea_13981_f2.tif

This article is protected by copyright. All rights reserved

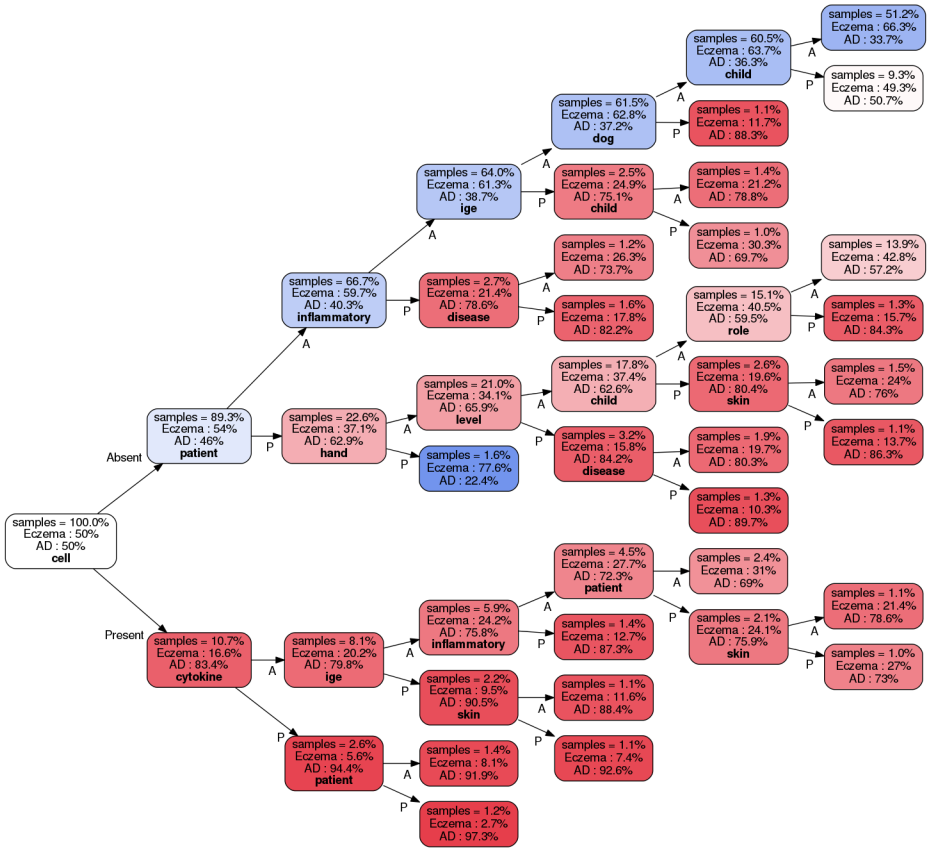
Accepted Article



cea_13981_f3.tif

This article is protected by copyright. All rights reserved

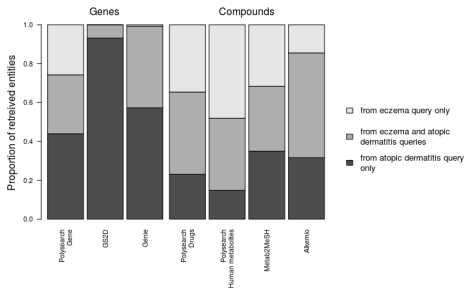
Accepted Article



cea_13981_f4.tif

This article is protected by copyright. All rights reserved

Accepted Article



cea_13981_f5.tif

This article is protected by copyright. All rights reserved

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

Clément Frainay, Yoann Pitarch, Sarah Filippi,

Marina Evangelou and Adnan Custovic

APPENDIX 1

A. MeSH co-occurences analysis

Table S1. Top 40 MeSH terms associated with eczema and AD.

Computed using Arrowsmith MeSH term Similarity Metrics tool : (http://arrowsmith.psych.uic.edu/arrowsmith_uic/mesh_pair_metrics.html). MeSH associated with eczema retrieved from a total of 8,599 articles. MeSH associated with atopic dermatitis retrieved from a total of 13,835 articles.

Associations with *eczema*				Associations with *dermatitis, atopic*			
MeSH Term	Article Count	Article Co-Occur	Article Odds Ratio	MeSH Term	Article Count	Article Co-Occur	Article Odds Ratio
Neurodermatitis	1412	444	102.9685	*Eczema*	8599	679	60.3341
Dermatitis, Atopic	13835	679	60.3341	*Food Hypersensitivity*	11737	1032	59.3172
Dermatitis, Contact	14292	804	59.3621	*Immunoglobulin E*	33625	2444	52.4149
Urticaria	9203	397	52.8488	*Pruritus*	8455	595	49.1492
Dermatitis, Occupational	7287	541	51.2699	*Dermatologic Agents*	8426	462	42.4945
Psoriasis	25918	863	47.5116	*Patch Tests*	8670	512	41.2637
Dermatitis	7469	360	44.3896	*Dermatitis, Contact*	14292	748	38.1594
Hand Dermatoses	6328	471	38.6129	*Emollients*	1307	205	36.9369
Patch Tests	8670	383	37.981	*Psoriasis*	25918	884	36.4146
Rhinitis, Allergic, Seasonal	12120	362	35.4832	*Malassezia*	1381	167	35.7755
Dermatitis, Seborrheic	2268	208	34.609	*Radioallergosorbent Test*	4139	258	34.799
Dermatitis, Allergic Contact	8286	274	31.5015	*Dermatitis, Seborrheic*	2268	186	33.2024
Pruritus	8455	266	29.2758	*Allergens*	30956	1342	32.3967
Ointments	10732	365	26.5262	*Antigens, Dermatophagoides*	2335	205	32.2733
Lichen Planus	4555	150	24.0462	*Tacrolimus*	12175	547	30.6855
Skin Diseases	47160	744	23.8584	*Kaposi Varicelliform Eruption*	445	102	30.2671
Hypersensitivity	36509	594	23.7771	*Milk Hypersensitivity*	1704	161	29.3581
Food Hypersensitivity	11737	272	22.8188	*Hypersensitivity, Immediate*	11392	492	29.0471
Prurigo	1020	122	22.2709	*Ointments*	10732	379	28.7644
Hypersensitivity, Immediate	11392	235	22.0161	*Water Loss, Insensible*	1781	145	28.5996
Rhinitis	8419	198	21.9853	*Respiratory Hypersensitivity*	8449	323	28.4481
Drug Eruptions	10895	236	21.6912	*Urticaria*	9203	433	28.334
FluocinoloneAcetonide	1183	99	21.1901	*Dermatitis, Occupational*	7287	283	27.9534
Skin Tests	30799	636	20.5175	*Chemokine CCL17*	456	86	27.6883
Administration, Topical	31428	385	20.0229	*Staphylococcal Skin Infections*	1755	157	27.3138
Dermatologic Agents	8426	166	19.9807	*Pyroglyphidae*	1058	110	27.0536
Asthma	100701	956	19.9316	*Skin Tests*	30799	1034	26.0874
Allergens	30956	407	19.2654	*Mites*	8983	334	25.28
Impetigo	1074	63	18.6722	*Administration, Topical*	31428	767	24.4081
Respiratory Sounds	6545	142	18.0112	*Rhinitis, Allergic, Perennial*	6802	327	23.942
Betamethasone	5085	165	17.9075	*Dermatitis, Allergic Contact*	8286	268	22.3296
Dermatitis, Exfoliative	2282	100	17.8253	*Intermediate Filament Proteins*	7089	208	22.2985
Emollients	1307	64	15.8967	*Rhinitis, Allergic, Seasonal*	12120	463	21.3502
Immunoglobulin E	33625	357	15.8667	*Intradermal Tests*	2907	163	21.0268
Acne Vulgaris	9259	158	15.4963	*Rhinitis*	8419	237	20.6842
Smallpox Vaccine	3490	81	15.4698	*Th2 Cells*	12646	358	20.1736
Pyoderma	2376	89	15.3926	*Conjunctivitis, Allergic*	2605	143	19.6213
Chromates	2183	63	15.3884	*Hand Dermatoses*	6328	222	19.3245
PhotosensitivityDisorders	5783	103	15.1382	*Langerhans Cells*	5047	160	19.0476
Rosacea	1897	53	14.8793	*Hypersensitivity*	36509	699	18.6848

B. Extracting biological entities using text mining tools

Each tool uses a different bio-entity reference set as well as different statistical methods to ensure the significance of the association. Consequently, the retrieved entities may vary between each software, and they can also exhibit large discrepancy regarding the number of entities retrieved, depending on the method used and the aim of the software. For each tool, we followed the default cut-off proposed by the developers, except for parameters limiting the number of returned results for display reasons. As the used solutions depend on external literature and biological databases, the results are expected to change following the updates and changes of the data used, in addition to the potential succeeding releases and modifications of their program.

Polysearch (http://polysearch.ca/) retrieves documents from local copies of several databases, including PubMed. The user query used to retrieve documents is expanded using its own thesaurus, automatically adding synonyms. Eczema and atopic dermatitis are not considered as synonyms in Polysearch's thesaurus. Sentences containing query terms and/or target entities (genes, metabolites, drugs...) are extracted. Each extracted sentence contributes to an overall relevance score of the target entity, according to several factors such as proximity between the query and target terms in the sentence. A list of negation terms is used to filter sentences such as "X is not induced by Y". A Z-score is computed for each gene or compounds from its relevance score. Entities without positive Z-score are filtered. We used queries "eczema" and "atopic dermatitis", selecting "Disease" as entry point. From the advanced search panel we keep default values, including automatic additions of synonyms, and PubMed as selected database. The maximum number of retrieved articles has been changed to "unlimited". All results and settings including synonyms are provided in the online repository.

GeneSet2Disease (http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=592) use Entrez Gene database to extract gene-related citations, from which the relationship with diseases are drawn using MeSH disease annotation of the corresponding PubMed entries. The relevance of the retrieved association is assessed from both a threshold of minimum number of disease related-citations for a given gene, and a statistical test of significance (Fisher exact test and Benjamini Hochberg multiple testing correction). For the *disease to gene* analysis, we used the default cut-off values: at least 5 disease-related citations for a gene to be considered, and a 0.05 threshold for corrected p-values. The number of rows to display has been changed from 25 to -1 (display all). All results and settings are provided in the online repository.

Metab2MeSH (http://metab2mesh.ncibi.org/) use substances and MeSH annotations of PubMed articles as well as compound name synonyms from PubChem entries to retrieve documents. Fisher exact test and Benjamini Hochberg multiple testing correction to assert the significance

of the association of substances and MeSH terms. We used the *search by MeSH term* functionality with exact match providing "eczema" and "dermatitis, atopic" as query. Metab2MeSH aims at providing an easy-to-use off-the-shelf solution and consequently provides little tuning parameters. Results provided in the online repository.

Génie (http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=6) and Alkemio (http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/?page_id=319) define a topic from a list of articles, through the form of a weighted list of discriminative words. A list of MeSH terms or a PubMed query can be used to retrieve articles, from which a sample is drawn to define the topic. Sets of abstract associated with biological entities are extracted from the NCBI database or chemical MeSH annotations of PubMed articles. The abstracts are classified as related or not to the topic based on the weighted list of discriminative words. Fisher exact test with FDR correction is then used to define the significance of the association between the entity and the topic. We build the topic using the MeSH search functionality, as the PubMed search option focus on recent articles only. Since the documents used are drawn from a random sampling, the results may vary due to different training sets used. The list of articles used and the topic model are provided in the online repository. We used the default user-defined cut-offs : P-value<0.01 for abstract selection and FDR<0.001. Results provided in the online repository.

C. Decision Tree for *eczema* and *atopic dermatitis* indexing

1 Import Documents from PubMed

In [ ]:
```
# coding: utf-8

import os
import requests
import xml.etree.ElementTree as et
```

1.1 Set access to PubMed webservices

In [ ]:
```
# pubmed webservice url for retreiving documents ids from query
url_search = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&retmode=xml&retmax=100000&term="
# pubmed webservice url for retreiving abstract from docuemnt id
url_fetch = "https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=text&rettype=xml&id="
```

1.2 Set PubMed queries

In [ ]:
```
#Pubmed queries
queries = {
    "eczema" : "(\"Eczema\"[MH] NOT \"Dermatitis, Atopic\"[MH]) AND (\"1945/01/01\"[PDat] : \"2017/31/12\"[PDat])",
    "dermatitis_atopic" : "(\"Dermatitis, Atopic\"[MH] NOT \"Eczema\"[MH]) AND (\"1945/01/01\"[PDat] : \"2017/31/12\"[PDat])"
}
```

1.3 Retrieve and parse response

In [ ]:
```
for label in queries :
    query = queries[label]
    pathData = "./data/raw/"+label
    if not os.path.isdir(pathData) :
        os.mkdir(pathData)


    # retrieve documents id related to query
    f = requests.get(url_search+query)
    #--parse response to extract article ids
    tree = et.fromstring(f.text)
    ids_list = tree.findall("./IdList/Id")
    print(query+"\t"+str(len(ids_list)))

    #--export PMIDs list
    with open("./data/"+label+"_PMIDs.tab", "w") as fout :
        for idPaper in ids_list :
            fout.write(str(idPaper.text)+"\n")

    # for each document returned by pubmed
    for idPaper in ids_list :

        #get pubmed entry xml content from pubmed identifier
        article = requests.get(url_fetch+idPaper.text)
        try :
            #get title from pubmed entry
            article_tree = et.fromstring(article.text)
            title = article_tree.find("./PubmedArticle/MedlineCitation/Article/ArticleTitle")

            #skip documents without title or undefined title
            if title is not None :
                document = title.text
                if "Not Available" not in document :

                    #get abstract content from pubmed entry
                    abstractObject = article_tree.find("./PubmedArticle/MedlineCitation/Article/Abstract/AbstractText")
                    #if abstract available, append to title in document
                    if not abstractObject is None :
                        document=document+" "+abstractObject.text
                    document = document.replace("[", "").replace("]", "") #remove brackets for translated titles
```

```
            #write extracted document
            with open(pathData+"/"+str(idPaper.text)+".txt", 'w') as fw :
                fw.write(document)
except Exception as e:
        print("error : unable to import data from PMID ",idPaper.text,"\n",str(e))
```

2 Document preprocessing

```
In [1]: import os
        import re
        import shutil


        import nltk
        from nltk.corpus import stopwords
        from nltk.tokenize import word_tokenize
        from nltk.stem import WordNetLemmatizer

        nltk.download('wordnet')
```

2.1 Set path to data

```
In [2]: input_dir = './data/raw/'
        output_dir ='./data/cleaned/'
```

2.2 Compile Regular expressions

Appart from removing special characters, some regex specific to PubMed documents has been defined : - removal of roman numbers, frequently used for book chapter titles - detection of missing content, (*[Not Available]*) - Resolving of scientific abbreviations

```
In [3]: # special character filter
        strip_special_chars = re.compile("[^\w_]+")
        # roman number filter
        strip_roman_numbers = re.compile("^[MDCLXVI]+[\s\.\n]")
        # missing data detector
        missing_data_detector = re.compile("\[Not Available\]\.")

        # atopic dermatitis abbreviation
        atopic_dermatitis_abbr = re.compile(r"\bAD\b")
```

2.3 Create Lemmatizer and stop words list

```
In [4]: lemmatizer = WordNetLemmatizer()

        def addStopwords(f) :
            res = []
            with open(f, "r") as fin :
                for l in fin :
                    res.append(l.strip())
            return res
```

2.4 Clean text documents

- remove roman number headings

- remove numerical values
- remove special characters
- remove words with length < 2 letters (include many units, usually not defined in stop words lists)
- remove stop words (wordnet)
- lower case
- tokenize
- lemmatize

processing specific to the study : - resolve atopic dermatitis abbreviation - resolve multiple-wording for atopic dermatitis and atopic eczema - lemmatize atopic dermatitis

```
In [5]: vocs = set()
        for label in os.listdir(input_dir) :
            path_in = input_dir+label+"/"
            path_out = output_dir+label+"/"
            if os.path.isdir(path_out):
                shutil.rmtree(path_out)
            if "DS_Store"not in path_out :
                os.mkdir(path_out)

                for fileName in os.listdir(path_in) :
                    if fileName.endswith(".txt"):
                        with open(path_in+fileName, "r") as fileContent :
                            content = fileContent.read().replace('\n',' ')

                            #remove roman number headings
                            content = re.sub(strip_roman_numbers, " ", content)

                            #remove numbers
                            #content = re.sub("(^|\W)[\d\s\.]+($|\W)", " ", content)
                            content = re.sub(r"\b\d+\b", " ", content)

                            #remove non alphabetical characters
                            content = re.sub(strip_special_chars, " ", content)

                            # switch to lower case
                            content = content.lower()

                            # QUERY SPECIFIC PREPROCESSING
                            # - resolve atopic dermatitis abbreviation
                            # - resolve multiple-wording for atopic dermatitis

                            content = re.sub(r"eczemas?(\b|\s)", "", content, flags=re.IGNORECASE)
                            content = re.sub("atopic", "", content, flags=re.IGNORECASE)
                            content = re.sub("dermatiti(s|des|ses)", "", content, flags=re.IGNORECASE)

                            for m in re.finditer(atopic_dermatitis_abbr, content):
                                content = re.sub(re.escape(m.group(0)), '', content)
```

```
                    #remove word with less than two letter
                    content = re.sub(r"\b\w{1,2}\b", '', content)

                    # remove stop words
                    stop_words = list(stopwords.words("english"))
                    #stop_words += addStopwords("stopwords.txt")
                    words = word_tokenize(content)
                    key_words = [word for word in words if word not in stop_words]

                    #lemmatize
                    lemmas = [lemmatizer.lemmatize(word) for word in key_words]

                    for l in lemmas :
                        vocs.add(l)

                    if lemmas :
                        with open(path_out+fileName,"w") as output :
                            output.write(" ".join(lemmas))

with open("vocab.txt", 'w') as fout :
    for w in vocs :
        fout.write(w+"\n")
```

3 Eczema and Atopic dermatitis articles classification

```
In [1]: import os
        import numpy as np
        from sklearn.feature_extraction.text import CountVectorizer
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import classification_report
        from sklearn import tree
        import graphviz
```

3.1 Data structure initialization

```
In [2]: tokenize = lambda doc: doc.lower().split(" ")
        dictLabel = {}
        labelId = 0
        path = "./data/cleaned/"

        labels = ["eczema", "AD"]
        labelIds = [0,1]
        XRawList = []
        yList = []
```

3.1.1 Function for creating balanced subsampled dataset

As the documents indexed under *dermatitis, atopic* outnumber those indexed under *eczema*, train set need to be balanced. We use down sampling, randomly selecting a number of documents from the majority class (*dermatitis, atopic*) equals to the number of documents from the minority class. As the training set will be different at each run, results might slightly vary.

```
In [3]: def balanced_subsample(x,y,subsample_size=1.0):

            class_xs = []
            min_elems = None

            for yi in np.unique(y):
                elems = x[(y == yi)]
                class_xs.append((yi, elems))
                if min_elems == None or elems.shape[0] < min_elems:
                    min_elems = elems.shape[0]

            use_elems = min_elems
            if subsample_size < 1:
                use_elems = int(min_elems*subsample_size)

            xs = []
            ys = []

            for ci,this_xs in class_xs:
```

```
            if len(this_xs) > use_elems:
            #if this_xs.shape[0] > use_elems:
                np.random.shuffle(this_xs)

            x_ = this_xs[:use_elems]
            y_ = np.empty(use_elems)
            y_.fill(ci)

            xs.append(x_)
            ys.append(y_)

        xs = np.concatenate(xs)
        ys = np.concatenate(ys)

        return xs,ys
```

3.2 Import documents

Documents correspond to preprocessed title and abstract retrieved from pubmed. Preprocessing include removal of stop words and convertion of words to lemma.

```
In [4]: for label in os.listdir(path) :

            if "Store" not in label and "ipynb_checkpoints" not in label:

                for txt in os.listdir(path+"/"+label):
                    if "ipynb_checkpoints" not in txt :
                        with open(path+"/"+label+"/"+txt) as finput :
                            XRawList.append(finput.read())
                            if "derma" in label :
                                yList.append(1)
                            else :
                                yList.append(0)


        y = np.asarray(yList)
        XRaw = np.asarray(XRawList)
```

3.3 Convert documents to feature vectors

Documents are represented as binary vectors, for which each position correpond to a word, and the value correspond to its presence or absence in the title or abstract. Each document has a known class definied by its MeSH indexing term in PubMed.

```
In [5]: # Vectorize documents
        # -- min_df : minimum frequency required to consider a word. Allow to reduce dimensions.
        # -- ngram_range : allow to consider ngrams (contiguous sequence of words) as feature.
        #    When set to [1,1], consider only single words for a simplier model.
```

```
        # -- binary : consider presence/absence of a word if true, number of occurences otherwise
        sklearn_count = CountVectorizer(min_df=10, tokenizer=tokenize,
                        ngram_range=[1,1], binary=True)
        X = sklearn_count.fit_transform(XRaw)
        # convert to dense matrix
        X = X.todense()

        # save vocabulary to translate features key back to words
        inv_map = {v: k for k, v in sklearn_count.vocabulary_.items()}
        feat_names = []
        with open("dict.txt", "w") as fout :
            for key in sorted(inv_map.keys()) :
                fout.write(str(key)+"\t"+inv_map[key]+"\n")
                feat_names.append(inv_map[key])
```

3.4 Learn Decision Tree

Build Decision Tree for *eczema* and *dermatitis, atopic* indexing classification, using CART algorithm with Gini impurity as split criterion. Maximum tree depth set to 6 and minimum proportion of total sample in a leaf set to 1%, in order to avoid overfitting and keep a simple model for ease of interpretation. The test set represent 20% of the whole dataset.

In [6]:
```
        # Randomly split dataset into train et test set
        X_train, X_test, y_train, y_test = train_test_split( X, y,
        test_size = 0.2, stratify=y, random_state = 42)

        # Balance trainset to avoid bias toward majority class
        X_train, y_train =  balanced_subsample(X,y,subsample_size=1.0)

        # Perform decision tree learning
        # -- criterion = Measure of the quality of a split, used to decide at each step
        #    which feature is best to split the data
        # -- max_depth = Maximum depth of the tree.
        # -- min_samples_leaf = The minimum proportion of total sample in a leaf.
        #    Below this value, no further split are performed. Prevent over-fitting
        clf = tree.DecisionTreeClassifier(criterion="gini", max_depth=6, min_samples_leaf=0.01)
        clf.fit(X_train,y_train)
```

Out[6]:
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=6,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=0.01, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                       splitter='best')
```

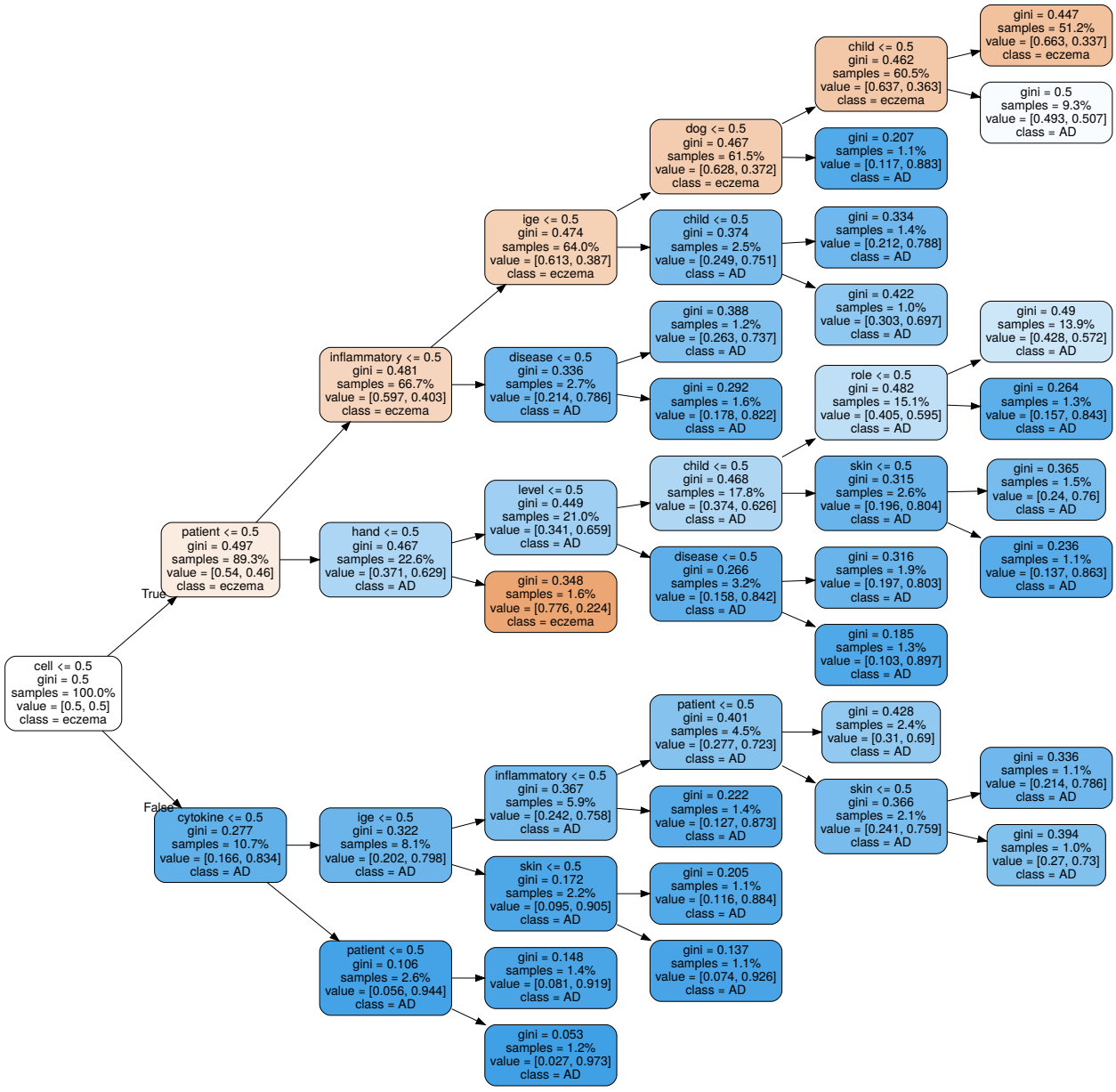
3.4.1 Create tree visualization

In [7]:
```
        #export tree vizualisation
        with open('./viz/tree.dot', "w") as f :
```

```
        f = tree.export_graphviz(clf,out_file=f,
                                 proportion=True,
                                 feature_names=feat_names,
                                 class_names=labels,
                                 filled=True,
                                 rounded=True,
                                 rotate=True)

    with open('./viz/tree.dot', "r") as f:
        dot_graph = f.read()
    graphviz.Source(dot_graph)
```

Out[7]:



3.5 Evaluation

Performance evaluated using F1 score, accuracy, specificity and ROC AUC.

```
In [8]: from sklearn.metrics import f1_score
        # apply classification model to test set
        y_pred = clf.predict(X_test)

        print(classification_report(y_test, y_pred,target_names=labels))

        # compute f1 score
        score = f1_score(y_test, y_pred, average='weighted')
        print("F1-score = ",(score*100))

        # compute ROC curve
        from sklearn.metrics import roc_curve, roc_auc_score
        import matplotlib.pyplot as plt

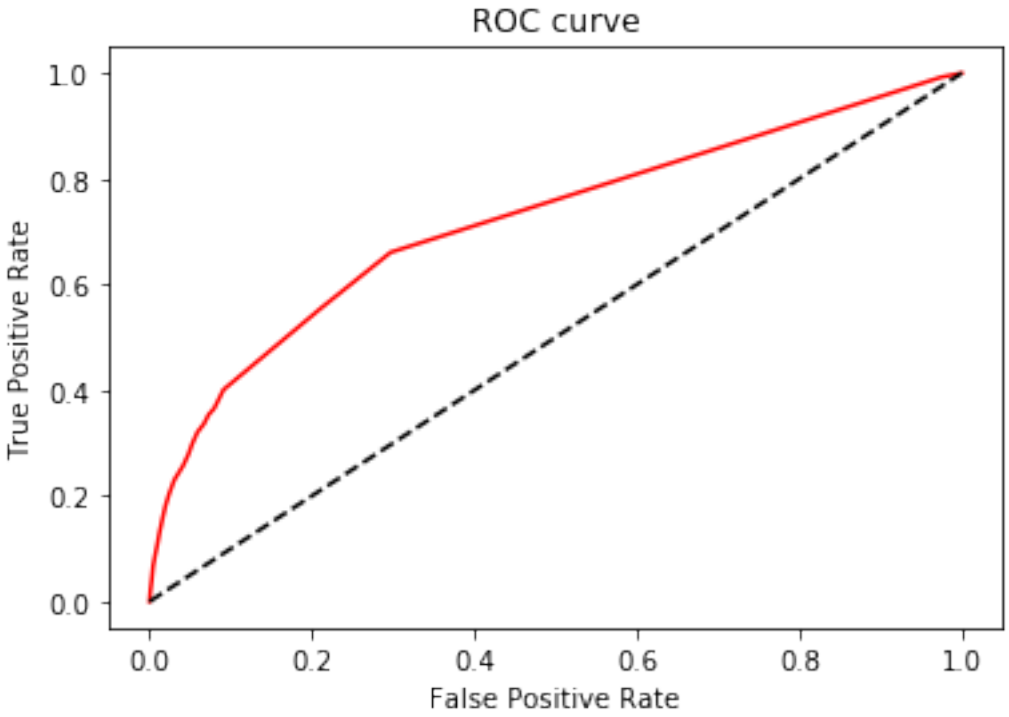
        y_prob = clf.predict_proba(X_test)
        fpr, tpr, _ = roc_curve(y_test, y_prob[:, 1],pos_label=1)
        plt.plot(fpr, tpr,'r-')
        plt.plot([0, 1], [0, 1],'k--')
        plt.title('ROC curve')
        plt.ylabel('True Positive Rate')
        plt.xlabel('False Positive Rate')
        plt.show()
        # print AUC score
        print("ROC AUC =",roc_auc_score(y_test, y_prob[:, 1]))
```

```
             precision    recall  f1-score   support

     eczema       0.53      0.70      0.61      1832
         AD       0.80      0.66      0.72      3309

avg / total       0.71      0.68      0.68      5141

F1-score =  68.2510681142661
```




```
ROC AUC = 0.7185735343993687
```