Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature

Thierry Hamon[a], Martin Graña[b], Víctor Raggio[b], Natalia Grabar[c,d], Hugo Naya[b]

[a] *LIMBIO (EA3969), UFR SMBH Léonard de Vinci, Université Paris 13, 74, rue Marcel Cachin, 93017 Bobigny France*
[b] *Unidad de Bioinformática, Institut Pasteur de Montevideo, Mataojo 2020, Montevideo 11400, Uruguay*
[c] *Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMRS 872, Paris, F-75006; Université Paris Descartes, UMRS 872, Paris, F-75006; INSERM, U872, Paris, F-75006 France*
[d] *HEGP AP-HP, 20 rue Leblanc, Paris, F-75015 France*

Abstract

Risk factors discovery and prevention is an active research field within the biomedical domain. Despite abundant existing information on risk factors, as found in bibliographical databases or on several websites, accessing this information may be difficult. Methods from Natural Language Processing and Information Extraction can be helpful to access it more easily. Specifically, we show a procedure for analyzing massive amounts of scientific literature and for detecting linguistically marked associations between pathologies and risk factors. This approach allowed us to extract over 22,000 risk factors and associated pathologies. The performed evaluations pointed out that (1) over 88% of risk factors for coronary heart disease are correct, (2) associated pathologies, when they could be compared to MeSH indexing, are correct in about 70%, and (3) in existing terminologies links between risk factors and their pathologies are seldom recorded.

Keywords:

Natural language processing, Semantics, Medical Informatics, Risk factors, Public Health Informatics, Terminology.

Introduction

Risk factors jointly refer to behaviors, environmental conditions, diseases or genetic backgrounds – considered in an ample way – that actually increase people's chance of manifesting a given disease. Discovering risk factors to design prevention strategies is an important biomedical challenge. Indeed, it is an active research field, with major contributors coming from biology, epidemiology and public health programs. Within these programs, research proposes statistical studies of large populations in order to reveal risk factors for many health conditions and pathologies (*i.e.* cardiovascular diseases [1, 2], hypertension [3], atherosclerosis [4, 5], cancers [6, 7], mortality within older people population [8, 9, 10], to cite but a few). Despite this extensive activity, traditional risk factors for coronary heart disease for instance actually account for only 50% of the risk [11]. For biomedical professionals, this research assumes pushing back the frontiers of medicine and biology. Learning about risk factors not only benefits preventive healthcare; it may directly impact on patients. Indeed, accumulated and organized clinical knowledge may help counteracting disease progression, hence improving life quality in patients. The last few years have seen the proliferation of websites intended to centralize and organize widely scattered medical and healthcare information. A major concern of health information collected from the Internet, is the very quality of the data [12, 13, 14]. The Medline database [15] provides biomedical researchers/practitioners with instant access to extensive and high quality scientific literature. Yet, the huge amount of indexed peer-reviewed articles, currently over 18 million, hinders reliably sampling and selecting all relevant papers. Our central goal is to exploit the Medline repository and to extract risk factors and their associations to health conditions.

Background

Currently, genetic and medical information in bioscience papers is often exploited in a manual, expert fashion. Hence, increasing attention is being paid to the automation of natural language processing (NLP) and information extraction (IE) in biological and medical realms, as testified by the BioCreAtIvE[1] and TREC Genomics[2] challenges dedicated to extraction of gene and protein names, their interactions and functions from biomedical literature; or by I2B2 challenges[3] dedicated to extraction of various kinds of information (i.e., smoking status, signs and symptoms, medication) from clinical records. However, to our knowledge, little work has focused on risk factor extraction.

As a matter of fact, existing approaches for risk factors study are mainly issued from the data mining area. Thus, an important issue faced by these works is the large number of variables

[1] `biocreative.sourceforge.net`
[2] `ir.ohsu.edu/genomics`
[3] `www.i2b2.org/NLP`

to be considered [10]: indeed, current state-of-the-art statistical algorithms are often unable to manage them. For example, data mining methods have been applied to ICD-9 codes to predict groups of people with similar risk factors [16], or to data from insurance companies to get estimators of claim costs [17]. Let us mention also a KDD challenge[4] held in 2004 which addressed data mining in biomedicine, albeit confined to social subgroups of patients, the main goal being to identify atherosclerosis risk factors (and their combinations) and to monitor the evolution of these risk factors as well as their impacts [18].

A rare work related to the processing of narrative biomedical literature [19] combines manual and automatic meta-analysis-like methods for extracting facts on risk factors related to breast cancer. Findings are consistent with published studies: there is a positive association with alcohol consumption, but a negative association with former smoking.

In this paper, we attempt taking a further step from the above-mentioned approaches. Our objectives are massive extraction of risk factors from available literature and establishing their relation(s) to the corresponding pathologies. Also, this will allow providing comprehensive information on risk factors, which seems to be a missing resource currently.

Material and Methods

Material

Bibliographical database Medline

Our rough working material is Medline database, which currently hosts over 18 million citations - mainly from life sciences and biomedical research. We exploit the totality of these data, focusing on titles, abstracts and *MeSH* indexing within each citation.

Bibliographical database Medline

MeSH [20] thesaurus has been created for information storage and retrieval, and is currently used for indexing Medline database. Its descriptors are organized in a hierarchical structure, whose most general level are very broad headings such as A *Anatomy*, C *Diseases*, L *Information Science*, F *Psychiatry and Psychology*, etc., with 16 such headings in total. Within *MeSH*, we exploit headings and their notations, and thus can rely on their semantic types.

Snomed CT

Snomed CT [21] is another terminological resource of the biomedical area. The goal of the *Snomed CT* nomenclature is to provide a conceptual basis for organizing and, more particularly, for exchanging clinical data. This is a multiaxial terminology. Its terms are organized within 15 hierarchies, such as: *Clinical finding/disorder*, *Procedure/intervention*, *Environment or geographical location, Social context, Organism, Substance, Pharmaceutical/biologic product,* etc. *Snomed CT* terms are structured within a dense network of semantic relationships belonging to synonymy, hierarchy, or transversal relationchips. Among this last category, we distinguished three more precise relationships, which in our opinion may be linked with the risk factor notion: has causative agent, due to and associated with. Thus, associated with represents a clinically relevant association between concepts without either asserting or excluding a causal or sequential relationship between the two, while due to is used to relate a clinical finding directly to its cause. In this example from *Snomed CT*: *acute pancreatitis due to infection* is a *acute pancreatitis* due to *infectious disease*, infectious disease belongs to the causative agent axis and allows to identify a direct cause of a disease. Other *Snomed CT* examples state that: (1) *bacterial endocarditis* has causative agent *bacterium*; and that *fentanyl allergy* has causative agent *fentanyl*. Within *Snomed CT* we exploit these three relationships which, as far as we know, may correspond to the only resource providing an explicit and controlled information on risk factors and associated health conditions.

Methods

Bibliographical database Medline

The aim of the preprocessing step is to annotate Medline citations with linguistic information and to prepare the step of information extraction task (identification of risk factors, of pathologies and of relations between them). We use the Ogmios platform [22], suitable for the processing of large amounts of data and tunable to specialized areas. The following tasks are performed through the platform: segmentation into words and sentences, POS-tagging and lemmatization with the Genia tagger [23], term extraction and recognition[5]. These tasks, and especially the term extraction task, are performed in order to correctly tokenize and identify textual units relevant to the searched information. Thus, the next step of information extraction is performed on POS-tagged, lemmatized and term-tagged text.

Information extraction

The proposed method is designed for accessing several types of information within Medline citations. First, we aim at detecting and extracting risk factors, and as a matter of fact, this implies dealing with specific syntactic structures, such as coordination and enumeration. In the following examples, risk factors are underlined, related pathologies and health conditions are in bold, and pattern-related elements in normal font letters:

1. Risk factors for ***survival*** were age and severity of aortic stenosis ... (PMID 8705769)
2. ...a high intake of calcium and phosphorus is a risk factor for the ***development of metabolic acidosis***. (PMID 1435825)
3. ...had more than one of the common risk factors for ***cerebrovascular accidents***, including hypertension, advanced age, hyperfibrinogenemia, diabetes mellitus, and past history of cerebrovascular accident. (PMID 1560589)

We process such syntactic structures with specifically designed syntactic patterns, where the trigger is a mention of *risk*

[4] lisp.vse.cz/challenge

[5] search.cpan.org/~thhamon/Lingua-YaTeA

factor and the enumeration sequence (punctuation and coordination conjunctions).

Another aspect of the information extraction task consists into establishing an explicit link between risk factors and the corresponding pathologies and health conditions. We propose to use two sources for accessing such information:

- Information extracted from abstracts and titles. In this case, this information is accessed through another set of dedicated lexico-syntactic patterns, such as *risk factors for* in the previous examples.
- *MeSH* descriptors provided by the Medline indexing. In this case, we analyze all the descriptors associated to a given citation, we then match them to the *MeSH* thesaurus and select only those which belong to the heading of diseases *C*.

Evaluation

We have to evaluate various aspects of the obtained results. Our main concern is the quality of extracted information for both risk factors and associated pathologies. We propose to tackle this as follows: (1) For a given pathology, we evaluate quality and exhaustiveness of the extracted risk factors. (2) As, for certain risk factors, we have at our disposal the associated pathologies provided by two sources (information extraction and *MeSH* indexing), we propose to compare these two pathology-related data. (3) Finally, we take advantage of the causal and associative relations encoded within *Snomed CT* and compare them with the information extracted from Medline titles and abstracts. We compute the precision, i.e. ratio of correct extractions among all the results relevant to a given evaluation. All these evaluations are performed manually, as no dedicated and comprehensive gold standard is available.

Results and Discussion

Building and preparing the material

Within the Medline database, we selected those citations which contain singular or plural forms of the terms *risk factor* and *factor of risk* in abstracts or titles. This allows to reduce the whole Medline material to a reliable and homogenous subset of citations. The resulting corpus contains 187,544 citations (over 42 M word occurrences). This corpus of Medline citations have been processed through the Ogmios platform. *Snomed CT* tables were accessed through the UMLS resource [24] version 2008AB, and we could extract 154,130 pairs of registered relations between a pathology and its causative agent or another pathology. 92,807 of these relations are provided by has causative agent, 25,309 by due to and 36,134 by associated with relationships. 120 of these relations are provided by more then one relationship within *Snomed CT*.

Extraction of information on risk factors and pathologies

Patterns for information extraction from abstracts and titles were built manually on few positive examples, and then generalized and applied on the whole set of data. We distinguish three kinds of patterns (in the examples below, <NP-RF> indicates noun phrases corresponding to risk factors, <NP-P> to pathologies, ? and * for optional and recurrent elements):

1. Patterns for extracting risk factors and pathologies (n=5). This example of pattern: <NP-RF> as a risk factor for <NP-P> allows to discriminate this sentence: *Hypocalcemia at parturition as a risk factor for left displacement of the abomasum in dairy cows,* and to propose that *hypocalcemia* is a risk factor of *displacement of the abomasum*.
2. Patterns for extracting risk factors (n=12), among which we have a pattern for enumeration: (other)? risk factors? including <NP-RF>(, <NP-RF>)* ((,)? and <NP-RF>)? . From the sentence *The relationships between impaired fasting glucose, other risk factors including blood pressure, and mortality have never been clearly investigated* it extracts that *blood pressure* and *mortality* are risk factors.
3. Patterns for extracting pathologies (n=3), among which (potential)? risk factors? for <NP-P>, which detects in the sentence *A risk factor for poor pregnancy outcome, a population-based screening study* the health condition *poor pregnancy outcome*.

In order to get complete information on associations between risk factors and pathologies, elements extracted by patterns (2) and (3) are combined. In this way, we assume that each citation corresponds to a semantically coherent unit and that elements from its different sentences are strongly related between them. These patterns allowed us to extract information from 10,445 PMIDs. Pattern (1) extracted 313 pairs {*risk factor*; *pathology*}. Combination of patterns (2) and (3) provided 15,398 pairs more. Finally, 5,873 risk factors extracted by pattern (2) could not be associated with any pathology within abstract or title. *MeSH* indexing was analyzed and allowed us to extract 5,106 different pathologies and health conditions within the axis C related to diseases. Exploitation of the proposed approach generates triplets {*risk factor*, $pathology_{text}$, $pathology_{MeSH}$}, where the first element is always informed, the two other elements may remain empty.

We extracted 21,584 triplets, among which 17,620 pairs (14,895 of which are unique) are provided only by information extraction patterns, while 5,717 pairs (4,412 of which are unique) contain *MeSH* descriptors as pathology.

Evaluation

Analysis of the extracted risk factors for coronary heart disease

Coronary heart disease (CHD) is the most common form of disease affecting the heart and is an important cause of premature death all around the world. A medical doctor performed a qualitative evaluation of results on this disease. This evaluation appears to be encouraging: among 1,102 extractions, only 128 (11.62%) are rejected, while the remaining set is considered to provide helpful information. First of all, well known risk factors (such as *hypertension*, *smoking*, *diabetes*, *age*, *obesity*, *hypercholesterolemia*, *hyperlipidemia*, *family history*) are frequently detected in the literature and extracted. Amusingly, *work* was detected to be a risk factor for CHD in the following sentence: *Passive smoking at work as a risk factor* *for coronary heart disease in Chinese women who have never smoked*. Obviously, this is an error due to an insufficient analysis of syntactic dependencies (the right risk factor is *passive smoking*). Such chunking and segmentation problems at sentence, term or word levels can appear but correspond usually to a minor problem. Another positive aspect of the method is that for several risk factors, it detects also synonyms: {*smoking; cigarette smoking; smoking history; importance of total life consumption of cigarettes*}, {*hyperhomocysteinemia; hyperhomocysteinaemia; homocysteine; plasma homocysteine*}.

Comparison between MeSH-indexed and extracted pathologies

In 291 cases, our approach generated a complete triplet {*risk factor, pathology*$_{text}$*, pathology*$_{MeSH}$}. In order to evaluate precision of the extracted pathologies, we propose to compare them with pathologies provided by *MeSH* indexing. This evaluation has been performed by a computer scientist and pointed out the following cases (in the given examples, pathologies are given in this order {*pathology*$_{text}$*; pathology*$_{MeSH}$}): 42 extracted pathologies are identical to *MeSH* indexing, 32 are their synonyms ({*breast cancer; breast neoplasms*}, {*coronary artery disease; coronary disease*}, {*cataractogenesis; cataract*}, {*postsurgical pain; pain, postoperative*}), 28 are lexically included ({*alzheimer; alzheimer disease*}, {*unsuspected anaphylaxis; anaphylaxis*}, {*hemorrhagic stroke; stroke*}, {*wound infection; surgical wound infection*}), 101 have a close semantic relation ({*poor pregnancy outcome; fetal growth retardation*}, {*development of alcohol disorders; alcoholism*}, {*stroke; cerebrovascular disorders*}, {*osteoporosis; bone diseases, metabolic*}, {*central retinal vein occlusion; vision disorders*}), 7 have a broad semantic relation ({*tardive dyskinesia; dyskinesia, drug-induced*}, {*squamous cell carcinoma of the skin; carcinoma, squamous cell*}) and 91 are not related semantically. Thus, among the set of 291 generated triplets, only 91 extracted pathologies (31%) appears to be not relevant, while nearly 70% are identical, or have a close or broad semantic relation with the *MeSH*-indexed pathologies.

Comparison of extracted risk factors with three **Snomed CT** *associative relations*

The evaluation question is whether relations extracted from abstracts and titles are already registered within *Snomed CT* and related through three causative relationships: has causative agent, due to and associated with. In order to analyze this aspect, we looked for those *MeSH*-indexed pathologies which are also involved in these three *Snomed CT* relationships. This evaluation is performed by a computer scientist. We obtain a set of 22,730 propositions, related to 168 various pathologies. We analyzed 20 pathologies (3,100 extractions, about 25% of the whole set), such as: *acquired immunodeficiency syndrome*, *kidney diseases*, *heart diseases*, *alcoholic intoxication*, *epilepsy*, and *cytomegalovirus infections*. Only 19 extractions (0.6%) were considered as already recorded within *Snomed CT* or very close to the recorded relations. For instance, within the sentence:

> *...how patients with abundant alcohol consumption as a risk factor develop the chronic alcohol abuse episode of care...* (PMID 10414608)

abundant alcohol consumption was extracted as risk factor for *alcoholism* C0001973, which is very close to the relation registered in *Snomed CT*: *drinking alcohol* (C0589068) has causative agent *alcoholism* (C0001973). Other comparable extractions: {*asbestos fibres* (C0003947); *asbestosis* (C0003949)} in *Snomed CT*, while in citations we obtain {*asbestos exposure*; *asbestosis* (C0003949)}; or {*cytomegalovirus group* (C0010825); *cytomegalovirus infections* (C0010823)} in *Snomed CT* and {*cytomegalovirus; cytomegalovirus infections* (C0010823)} in text. Remaining extractions share little common aspects with the analyzed *Snomed CT* associative relations. One reason is that the precision of these extractions is not perfect; but a more specific reason is that *Snomed CT* does not specifically record this type of information, although risk factors may occur among the *Snomed CT* relations. Thus, for *acquired immunodeficiency syndrome* we extracted several risk factors, i.e.:

> *bisexuality* (C0005639), *bisexual* (C0178515),
> *blood transfusion* (C0005841),
> *intravenous drug abuse* (C0086181), ...

which are *Snomed CT* concepts but have no associative relations to *acquired immunodeficiency syndrome*. The situation is similar with other illnesses: *family history*, *age*, *race*, *smoking*, *hyperlipidemia*, *diabetes*, *hypertension*, *sedentary life style*, *weight control*, *stress* and many others are extracted as risk factors for *heart diseases* but are not recorded as such in *Snomed CT*. This implies, not surprisingly, that the creation and maintenance of specific dedicated and comprehensive resource for risk factors, would be most welcome.

Conclusion and Perspectives

We presented an experience in extracting information linked to risk factors from Medline citations. The proposed approach relies on NLP and IE methods and is based on lexico-syntactic patterns. It allows extracting risk factors and the associated pathologies. We perform several types of evaluation by medical and biological experts and computer scientists. Evaluation of the precision of risk factors extracted for coronary heart disease shows that they cover a large range of risks, and that only 11.62% of them are incorrect, while the remaining 88.38% are correct. Comparison of the associated pathologies extracted from abstracts with pathologies provided by the *MeSH* indexing prove to be identical or semantically related in about 70%. These two evaluations are very positive. Finally, a comparison of pairs {*risk factor*; *pathology*} extracted from citations and those proposed by associative relationships within *Snomed CT* allowed us to observe that *Snomed CT* is not dedicated to the recording of this type of information, although some of the pathologies can be related to their risk factors in this terminological resource.

We have several perspectives for this work. Methodologically, we will design and apply other patterns for detecting and extracting information on risk factors. For instance, triggers like *predictor, precursor* are not taken into account currently. We also plan to apply other methods, *i.e.* machine learning and text mining. From a knowledge representation viewpoint, a more precise categorization of risk factors within homogenous groups will be performed. Thus, we can distinguish groups related to environmental, social, clinical, behavioral, and other risks. Besides, this categorization can even be mentioned in abstracts:

> *Demographic risk factors (age, sex, and ethnicity), clinical risk factors (diabetes mellitus, increased cholesterol, antihypertensive medications, history of congestive heart failure, myocardial infarction, hypertension, and neurological deficits), and behavioral risk factors (smoking and heavy drinking) were controlled for statistically.* (PMID 11973166)

Another perspective is related to characterizing the extracted information itself. Thus, some of the extracted elements may occur in modal or negative contexts, which reduces their reliability. Otherwise, geographical, demographic or other variables for risk factors exist. For instance, for a given pathology, common risk factors in North America or Europe may be of less (or no) relevance to other geographical areas. Such characterization may well deserve more focus.

References

[1] Gouni-Berthold I, Krone W, and Berthold H. Vitamin d and cardiovascular disease. Curr Vasc Pharmacol 2009;7(3):414–22.

[2] Perret-Guillaume C, Joly L, and Benetos A. Heart rate as a risk factor for cardiovascular disease. Prog Cardiovasc Dis 2009;52(1):6–10.

[3] Meneton P, Galan P, Bertrais S, et al. High plasma aldosterone and low renin predict blood pressure increase and hypertension in middle-aged caucasian populations. J Hum Hypertens 2008;22(8):550–8.

[4] Dentali F, Squizzato A, and Ageno W. The metabolic syndrome as a risk factor for venous and arterial thrombosis. Semin Thromb Hemost 2009;35(5):451–7.

[5] Portugal L, Fernandes L, and Alvarez-Leite J. Host cholesterol and inflammation as common key regulators of toxoplasmosis and artherosclerosis development. Expert Rev Anti Infect Ther 2009;7(7):807–19.

[6] Turnbull C and Hodgson S. Genetic predisposition to cancer. Clin Med 2005;5(5):491–8.

[7] Fleshner N and Lawrentschuk N. Risk of developing prostate cancer in the future: overview of prognostic biomarkers. Urology 2009;73(5):21–7.

[8] Glynn R, Field T, Rosner B, et al. Evidence for a positive linear relation between blood pressure and mortality in elderly people. Lancet 1995;345(8953):825–9.

[9] Bennett K. Low-level social engagement as a precursor of mortality among people in later life. Age Ageing 2002;31:165–8.

[10] Ahmad R and Bath PA. Identification of risk factors for 15-year mortality among community-dwelling older people using Cox regression and a genetic algorithm. Journal of Gerontology 2005;60(8):1052–8.

[11] Allen J. Genetics and cardiovascular disease. Nurs Clin North Am 2000;35(3):653–2.

[12] Boyer C, Baujard O, Baujard V, et al. Health on the net automated database of health and medical information. Int J Med Inform 1997;47(1-2):27–9.

[13] Darmoni S, Leroy J, Baudic F, et al. CISMeF: cataloque and index of french speaking health resources. In: Stud Health Technol Inform, 1999:493–6.

[14] Risk A and Dzenowagis J. Review of internet information quality initiatives. Journal of Medical Internet Research 2001;3(4).

[15] NLM . Medline: medical literature on-line. National Library of Medicine, Bethesda, Maryland, 2009. www.ncbi.nlm.nih.gov/sites/entrez.

[16] Cerrito P. Inside text mining. Health management technology 2004;25(3):28–31.

[17] Kolyshkina I and van Rooyen M. Text mining for insurance claim cost prediction. Springer-Verlag, 2006:192–202.

[18] Brisson L, Pasquier N, Hebert C, and Collard M. Hasar: mining sequential association rules for atherosclerosis risk factor analysis. In: PKDD 2004, Pisa Italy. 2004:14–25.

[19] Blake C. A text mining approach to enable detection of candidate risk factors. In: Medinfo, 2004:1528–.

[20] National Library of Medicine, Bethesda, Maryland. Medical Subject Headings, 2001. www.nlm.nih.gov/mesh/meshhome.html.

[21] Stearns M, Price C, Spackman K, and Wang A. Snomed clinical terms: overview of the development process and project status. In: AMIA, 2001:662–6.

[22] Hamon T, Nazarenko A, Poibeau T, Aubin S, and Derivière J. A robust linguistic platform for efficient and domain specific web content analysis. In: RIAO 2007, Pittsburgh, USA. 2007.

[23] Tsuruoka Y, Tateishi Y, Kim JD, et al. Developing a robust part-of-speech tagger for biomedical text. LNCS 2005;3746:382–92.

[24] NLM . UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland, 2007. www.nlm.nih.gov/research/umls/.

Address for correspondence

Thierry Hamon, LIMBIO (EA3969)
UFR SMBH Leonard de Vinci, Universite Paris 13
74, rue Marcel Cachin, 93017 Bobigny Cedex France[.5ex]
e-mail: thierry.hamon@univ-paris13.fr