Journal of Biomedical Informatics 70 (2017) 27–34

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews

George Karystianis [a,*], Kristina Thayer [b], Mary Wolfe [b], Guy Tsafnat [a]

[a] Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[b] National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA

ARTICLE INFO

Article history:
Received 12 December 2016
Revised 14 March 2017
Accepted 2 April 2017
Available online 25 April 2017

Keywords:
Text mining
Rule-based modelling
Dictionaries
Environmental health studies
Automation of systematic reviews

ABSTRACT

Introduction: Most data extraction efforts in epidemiology are focused on obtaining targeted information from clinical trials. In contrast, limited research has been conducted on the identification of information from observational studies, a major source for human evidence in many fields, including environmental health. The recognition of key epidemiological information (e.g., exposures) through text mining techniques can assist in the automation of systematic reviews and other evidence summaries.
Method: We designed and applied a knowledge-driven, rule-based approach to identify targeted information (study design, participant population, exposure, outcome, confounding factors, and the country where the study was conducted) from abstracts of epidemiological studies included in several systematic reviews of environmental health exposures. The rules were based on common syntactical patterns observed in text and are thus not specific to any systematic review. To validate the general applicability of our approach, we compared the data extracted using our approach versus hand curation for 35 epidemiological study abstracts manually selected for inclusion in two systematic reviews.
Results: The returned F-score, precision, and recall ranged from 70% to 98%, 81% to 100%, and 54% to 97%, respectively. The highest precision was observed for exposure, outcome and population (100%) while recall was best for exposure and study design with 97% and 89%, respectively. The lowest recall was observed for the population (54%), which also had the lowest F-score (70%).
Conclusion: The generated performance of our text-mining approach demonstrated encouraging results for the identification of targeted information from observational epidemiological study abstracts related to environmental exposures. We have demonstrated that rules based on generic syntactic patterns in one corpus can be applied to other observational study design by simple interchanging the dictionaries aiming to identify certain characteristics (i.e., outcomes, exposures). At the document level, the recognised information can assist in the selection and categorization of studies included in a systematic review.

1. Introduction

Observational epidemiological studies are a valuable information source of human evidence in many fields, including environmental health, health care, nutrition, public health, and the social sciences [1]. Unlike experimental research, observational studies do not rely in researcher-implemented interventions but are focused in understanding the etiology [1] and prognosis of exposures along with their relations to targeted outcomes by observing a population of interest. And this is where their strength relies; in the assessment of the outcome causality since they can provide important pieces of evidence for clinical decision making and potentially shorten the time required to link the evidence to in-development public health policies, and the implementation of population-based public health interventions [2]. Manually extracting information about observational studies from papers is a time-consuming task [2,3] that is nonetheless required in order to synthesize and compare the results of multiple studies in a systematic review [4,5]. Thus, there is a pressing need to develop technologies that can help automate the analysis of scientific literature, including the provision of quick access to information found in large volumes of documents [6].

Abbreviations: RCT, randomized clinical trial; NLP, Natural Language Processing.

* Corresponding author at: Australian Institute of Health Innovation, Centre for Health Informatics, Level 6, 75 Talavera Road, North Ryde, Sydney, NSW 2113, Australia.
E-mail address: george.karystianis@mq.edu.au (G. Karystianis).

http://dx.doi.org/10.1016/j.jbi.2017.04.004
1532-0464/

Natural Language Processing (NLP) has been used to recognise key biomedical information contained in clinical notes and biomedical journal articles [7–10]. The majority of related studies in epidemiological research involve the identification of key knowledge from clinical trials since they are considered the most informative study design to assess causality [1,2,11–17]. There is a wealth of health effects data in publications of observational human studies with a range of designs including retrospective and prospective cohort, case-control, and cross-sectional. Those data can help infer causality, reveal research trends regarding a particular disease, highlight susceptible populations to a targeted exposure, or point out areas for future research affecting existing hypotheses (i.e., confounding factors). Compared to reports of clinical trials, observational research is frequently unclear, lacks details, and is reported through less structured and more descriptive prose. This represents a challenge for text analytics research.

Biomedical text mining employs NLP techniques that can assist researchers in the recognition and integration of key knowledge from text, which can then be utilized to synthesize evidence from within and across studies and fuel further scientific research and exploration [18–20]. Its role is to help researchers make sense of large amounts of text by distilling information and extracting facts and facilitate the generation of hypotheses relevant to the user's information needs [6,21]. A number of studies have indicated the feasibility of text mining for the recognition and association of information relevant to a given health care problem in biomedical text [21–24].

Observational studies offer a variety of characteristics that could be potentially targeted for extraction with the most common being [25]:

- **Study design:** the implemented plan or protocol for the conduction of the study.
- **Population:** the number of individuals participating in the observational study including, if mentioned, demographic attributes (population size, related ethnicity, nationality, age group, gender).
- **Exposure:** the factor or entity that causes or may be associated with a change in the health condition or any other attribute of the participant population.
- **Outcome:** the consequence from the exposure in the population of interest.
- **Confounders:** a factor associated with the exposure of interest that may be a potential cause of the outcome of interest. Confounders can lead to bias that distort the magnitude of the relationship between two factors of interest.
- **Country:** the country where the epidemiological study was conducted.

Since epidemiology is a field in which studies follow a semi-structured reporting style, with its own specific "dictionary", we hypothesized that a knowledge-driven approach like NLP (i.e., rules that can identify targeted characteristics of interest) could provide an effective means to extract key information from text.

In this paper, we present the evaluation of a methodology that enables the recognition of key study characteristics from observational study abstracts that assess health impacts associated with environmental chemical exposures.

2. Materials and methods

2.1. Source data

We obtained epidemiological study documents in abstract form previously collected for five independent systematic review topics, assembled by the National Toxicology Program (NTP),[1] that are being evaluated in the application of systematic review methodology to environmental health topics. Data from hand curation of these studies were accessed via the HAWC [26] (Health Assessment Workspace Collaborative) software, a free and open-source data content system (https://hawcproject.org). We used five corpora of epidemiological studies with various environmental exposures and associated outcomes (Table 1) as access was provided for only these five: three corpora were used for the design of rules and the creation of dictionaries with one corpus utilized as the initial training set and two more as development sets focusing on the optimization of the rule-based methodology; finally, two more corpora previously unseen were used for the evaluation of our method.

2.2. Knowledge based system development

Our methodology involved the design and implementation of generic rules that enable the recognition of mentions of targeted elements in the right context in epidemiological study abstracts. Interchangeable dictionaries of targeted elements make the rules independent of specific study areas. For example, one dictionary includes all possible forms of perfluorooctanoic acid (PFOA) so that, when used with exposure extraction rules, information was extracted about exposures to PFOA. The rule and dictionary development process is summarized in Fig. 1 and explained further below. By changing the dictionary (and making it exposure specific), and not the rules, the system is able to identify exposures from other observational studies. In particular, we can apply the (same) set of rules to a different corpus of epidemiological study abstracts related with environmental exposures by supplying a dictionary that contains terms related with the exposure and outcomes of the corpus itself.

The rules are based on common lexical patterns (e.g., "***perfluorooctanoate (PFOA)*** [exposure] in relation to ***weight and size at birth*** [outcome]"; 'A total of ***428 women and their infants*** [population] were involved in the study") that suggest the presence of key elements observed in text. To be more specific, the lexical patterns use frozen lexical expressions as anchors for certain elements (e.g., "***study design:*** cross-sectional study") along with semantic classes that are identified through application of the manually crafted dictionaries. Verbs, noun phrases, and prepositions are included in the frozen expressions. However, no part-of-speech tagging is being used in the text mining pipeline. More than one of the lexical patterns may exist in a study abstract and refer to one or more characteristics; e.g., "*effect of X on Y*" (pattern for exposure/outcome), "*after controlling for X and Y*" (pattern for confounder).

2.2.1. Rules

In order to create the rules, we used General Architecture for Text Engineering (GATE) [32], a text mining framework for annotating and categorizing text with its own user-interface that enables the extraction of targeted data. GATE was chosen due to its support of rule-based, text-mining approaches and its effective graphical user interface. The observed lexical patterns in text were converted into rules following the GATE schema with the usage of regular expressions and the application of small vocabularies that contain synonymous words (e.g., verbs indicating the registration of the studied population for an exposure dose: "recruited", "enrolled", "registered"). GATE uses tokenization of words and symbols (e.g., "?", ",") and the grammatical nature of each token is

[1] NTP is an interagency program located at the National Institute of Environmental Health Sciences, part of the U.S. National Institutes of Health. NTP evaluates substances of public health concern through toxicology testing, research, and literature analysis activities (http://ntp.niehs.nih.gov).