Computers in Biology and Medicine 73 (2016) 173–185

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Text mining, a race against time? An attempt to quantify possible variations in text corpora of medical publications throughout the years

Mathias Wagner [a], Benjamin Vicinus [b,c], Sherieda T. Muthra [d,*], Tereza A. Richards [e], Roland Linder [f], Vilma Oliveira Frick [b], Andreas Groh [g], Claudia Rubie [b], Frank Weichert [h]

[a] Department of Pathology, University of Saarland, Homburg Saar Campus, Homburg Saar, Germany
[b] Department of General, Visceral, Vascular and Pediatric Surgery, University of Saarland, Homburg Saar Campus, Homburg Saar, Germany
[c] Institute of Virology, University of Saarland, Homburg Saar Campus, Homburg Saar, Germany
[d] Lombardi Comprehensive Cancer Center, Georgetown University, 37th & O St NW, Washington, DC 20057, United States of America
[e] The Medical Library, University of the West Indies, Mona, Kingston, Jamaica
[f] Institute of Medical Informatics, University of Luebeck, Luebeck, Germany
[g] Department of Mathematics, University of Saarland, Saarbrücken Campus, Saarbrücken, Germany
[h] Department of Computer Science VII, Technical University of Dortmund, Dortmund, Germany

ARTICLE INFO

Article history:
Received 3 December 2015
Received in revised form
19 March 2016
Accepted 21 March 2016

Keywords:
Nomenclature
Systems biology

ABSTRACT

Background: The continuous growth of medical sciences literature indicates the need for automated text analysis. Scientific writing which is neither unitary, transcending social situation nor defined by a timeless idea is subject to constant change as it develops in response to evolving knowledge, aims at different goals, and embodies different assumptions about nature and communication. The objective of this study was to evaluate whether publication dates should be considered when performing text mining.

Methods: A search of PUBMED for combined references to chemokine identifiers and particular cancer related terms was conducted to detect changes over the past 36 years. Text analyses were performed using freeware available from the World Wide Web. TOEFL Scores of territories hosting institutional affiliations as well as various readability indices were investigated. Further assessment was conducted using Principal Component Analysis. Laboratory examination was performed to evaluate the quality of attempts to extract content from the examined linguistic features.

Results: The PUBMED search yielded a total of 14,420 abstracts (3,190,219 words). The range of findings in laboratory experimentation were coherent with the variability of the results described in the analyzed body of literature. Increased concurrence of chemokine identifiers together with cancer related terms was found at the abstract and sentence level, whereas complexity of sentences remained fairly stable.

Conclusions: The findings of the present study indicate that concurrent references to chemokines and cancer increased over time whereas text complexity remained stable.

1. Introduction

The current biomedical domain is a repository for a tremendous body of knowledge available in the form of natural language text. The quantity of academic reports concerning the biological activities of chemokines is increasing at a high rate. This proliferation of specialized knowledge presents challenges to understanding the larger picture of chemokine mechanisms and function. In order to overcome the "analysis paralysis" associated with large quantities of data and the resulting information overflow, various computer-aided text mining systems have been developed to extract accumulated information directly from easily accessible public data repositories such as PubMed at high throughput scale [1].

Mining algorithms that use concepts of similarity or distance tend to search for the closest match to a query by ranking across database objects. They appear to be based on the assumption that there are no relevant time dependent changes of variables such as keyword co-appearance or readability score results of text corpora. Systematic studies focusing on possible changes in grammar and terminology over time are scarce when viewed in light of applications of mining algorithms to biomedical databases. Hence, it may be considered indeterminate whether the body of biomedical

* Corresponding author.
E-mail address: stm36@georgetown.edu (S.T. Muthra).

http://dx.doi.org/10.1016/j.compbiomed.2016.03.016
0010-4825/

literature conveys possibly relevant changes in readability and concurrent references to chemokines and cancer related terms over time. The present study was therefore conducted to evaluate whether there is substantial linguistic change associated with the relative date of abstract publication. Positive findings could be suggestive of a need for an adaptation of text mining algorithms to linguistic properties possibly characteristic for particular dates of publication.

It is historically wrong to treat the style of scientific writing as fixed and epistemologically neutral. The present study was conducted to assess whether scientific writing displays considerable readability score changes over time along with the co-mentioning of chemokines and cancer related terms.

Chemokines are members of a superfamily of chemo-attracting, cytokine-like proteins that constitute regulators of cell trafficking and adhesion. While for many years the main function of chemokines focused on inflammation and wound repair, other functions became increasingly obvious in the last decade. Some chemokines since then appear to be involved in mediating tumor metastasis in various cancer entities, hereby constituting important therapeutic targets. Histology shows that increased rates of co-occurrence of chemokines and cancer-related terms does not necessarily mean that chemokines are found in tumor cells. The main motive for putting the focus on chemokines was the fact that a unified nomenclature has been repeatedly proposed by an official committee, the *IUIS/WHO Subcommittee on Chemokine Nomenclature* and that this nomenclature appears to be commonly used. Selecting chemokines was therefore assumed to reduce effects of possibly ambiguous nomenclature. Cancer was chosen as it certainly is a substantial public health concern while chemokines were selected due to the considerable prevalence of acronyms and abbreviations, orthographic and lexical variants, homonyms (*i.e.*, words that have multiple unrelated meanings) and aliases (*i.e.*, alternative names by which a particular chemokine is known).

2. Materials and methods

The authors focused on using freeware. Abstracts with combined mentions of chemokine identifiers and particular search terms were examined. Each of the latter had at least one association with solid (*i.e.*, non-hematological) malignancy, possibly suggesting or negating interactions between individual chemokines and entities known to present more or less malignant biological behavior.

2.1. Raw data collection

A slightly modified version of the true positives of the *Top 25 List of Malignancy Mentions* (provided by the automatic malignancy mention extractor *MTAG*) was interactively created by removing organ-specific constituents from composite expressions to support universality [2]. The term "neoplasm", known to be "neutral" with regard to benign or cancerous behavior, was indirectly included by application of untagged cancer-related terms leading to a match against a *Medical Subject Headings* (*MESH*) translation table. The possible search terms "leukemia" and "lymphoma" were omitted as these classifiers were not considered to usually label solid tumors. The words "carcinogenesis" and "cancerogenesis" were added to the list of the remaining true positives to include etiological aspects. The terms of this now modified list were coupled with the chemokine identifiers repeatedly proposed by the *IUIS/WHO Subcommittee on Chemokine Nomenclature* [3–5] to search the *PUBMED* database on July 16th 2010. Search strategies such as "CCL1 AND (cancer OR carcinoma OR sarcoma OR metastasis OR carcinogenesis OR cancerogenesis)" were then adapted to all human chemokines of the CC and CXC type (chemokine identifiers were considered to be based on cysteine subclass roots, followed by "L" for "ligand" while the numbers were identical to the numeric identifiers used in the corresponding gene nomenclature).

The *MTAG*'s *Top 25* word "melanoma" was arbitrarily selected to be excluded from the search terms to help estimate the error of the second kind (*syn.*: type II error). This was exemplified by the number of melanoma-related findings that were produced by the aforementioned search strategy without the name of this solid tumor being among the search terms. In this context it is noteworthy that the term "melanoma growth stimulating factor alpha" was known to be an alias representing the chemokine CXCL1. Additional control datasets were generated by posing the query "chemokine AND (cancer OR carcinoma OR sarcoma OR metastasis OR carcinogenesis OR cancerogenesis)" and "microRNA AND (cancer OR carcinoma OR sarcoma OR metastasis OR carcinogenesis OR cancerogenesis)" to *PUBMED* with the terms "chemokine" and "microRNA" serving as controls. The *MESH* analysis inclusion feature was not deactivated. This search was also performed with untagged terms to induce a match against the *MESH* translation table. A 3-year interval was selected for analysis as this range was best suited for the analysis tools we utilized. The search was limited to entries with available abstracts. Findings were differentially analyzed for all triennials from January 1st, 1974, to December 31st, 2009 (*Publication Date*; *Pubdate*). It is noteworthy that the search was not designed to ensure guaranteed retrieval of the first mention of any chemokine alluded to a cancer-related term. Human curators manually isolated the abstracts from the *Extensible Markup Language* (*XML*) format provided by *PUBMED* to support further semi-automated analyses.

Abstracts of Open Access publications and traditional journal collections were treated equally as their linguistic structures were considered to be indistinguishable from each other [6]. Abstracts were assumed to linguistically represent full text corpora as word occurrence, lexical and intellectual density in abstracts have been shown to be comparable to those of full texts [7]. Text analyses introduced in the present study was therefore restricted to abstracts and (to a certain extent) institutional affiliations.

2.2. Readability analyses

The authors of the present study hypothesized that alias usage and publication date might be associated with circumlocution. Several parameters were therefore analyzed to estimate the readability of the abstracts: the number of words, characters, and sentences, the average number of characters or syllables per word, and the average number of words per sentence. In addition, the *Flesch Reading Ease* (*FRE*), the *Flesch-Kincaid Grade* (*FKG*) *Level*, the *Gunning Frequency of Gobbledygook* (*FOG*) *Index*, the 1975 *Coleman Liau Index* (*CLI*), and the *Automated Readability Index* (*ARI*), were assessed.

2.3. Similarity analysis

The current body of literature was examined to demonstrate possible associations between the results of the aforementioned linguistic analyses and the current level of knowledge on chemokines in the context of cancer. A rather unorthodox strategy was performed to obtain numeric data on texts. Publication dates were arbitrarily not considered in this part of the study.

As language complexity can be quantified [8], text passages were designed at the sub-sentence level (representing little grammatical complexity) using precise vocabulary to reduce putative bias associated with readability. Series of word aggregates containing a chemokine name and a cancer-related term were