




# Special Issue on the Curative Power of Medical Data

Daniela Gifu<sup>1,2,3,\*</sup> , Diana Trandabăț<sup>2</sup> , Kevin Cohen<sup>4</sup> and Jingbo Xia<sup>5</sup> 

<sup>1</sup> Institute of Computer Science, Romanian Academy-Iasi branch, Iasi 700481, Romania

<sup>2</sup> Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Iasi 700483, Romania; dtrandabat@info.uaic.ro

<sup>3</sup> Cognos Business Consulting S.R.L., 7, Iuliu Maniu Blvd, Bucharest 061072, Romania

<sup>4</sup> Computational Bioscience Program, University of Colorado School of Medicine, Aurora, CO 80045, USA; kevin.cohen@gmail.com

<sup>5</sup> Department of Biostatistics, Huazhong Agricultural University, Wuhan 430070, China; xjb@mail.hzau.edu.cn

\* Correspondence: daniela.gifu@info.uaic.ro or daniela.gifu@iit.academiaromana-is.ro or daniela.gifu73@gmail.com; Tel.: +40-742-050-673

Received: 12 June 2019; Accepted: 12 June 2019; Published: 14 June 2019



**Abstract:** With the massive amounts of medical data made available online, language technologies have proven to be indispensable in processing biomedical and molecular biology literature, health data or patient records. With huge amount of reports, evaluating their impact has long ceased to be a trivial task. Linking the contents of these documents to each other, as well as to specialized ontologies, could enable access to and the discovery of structured clinical information and could foster a major leap in natural language processing and in health research. The aim of this Special Issue, “Curative Power of Medical Data” in *Data*, is to gather innovative approaches for the exploitation of biomedical data using semantic web technologies and linked data by developing a community involvement in biomedical research. This Special Issue contains four surveys, which include a wide range of topics, from the analysis of biomedical articles writing style, to automatically generating tests from medical references, constructing a Gold standard biomedical corpus or the visualization of biomedical data.

**Keywords:** biomedical research; medical text mining; health data analyzing

## 1. Introduction

The interest in biomedical data, along with the continuous development of various qualitative and quantitative text analyses tools, made language technologies a natural choice for analyzing the evolution of the scientific life.

Biomedical Text Mining (BTM), known also as Health Informatics (HI), is a discipline situated at the intersection of computer science, medicine care, information systems and linguistics. It focuses on the resources, methods and applications needed to optimize the acquisition, storage, recovery and mining of information in biomedicine. Additionally, BTM uses sophisticated predictive models to understand, identify and extract concepts from a large collection of scientific texts in medicine, biology, biophysics, chemistry, etc., with the aim to discover knowledge that can add value to medical research [1].

In this context, a wide range of language resources has been developed, including complex lexicons, thesauri and ontologies that cover the entire spectrum of clinical concepts. In order to provide a uniform conceptual understanding, some authors have described a terminological and typology system [2,3].

Innovative approaches for the understanding of biomedical information using semantic web technologies and linked data have started to be investigated by bringing together practitioners, researchers, and scholars to share examples, use cases, theories and analyses of biomedical data. A

suitable forum for this was the MEDA Workshop on the Curative Power of Medical Data<sup>1</sup>, the starting point for this Special Issue, aimed at consolidating an internationally appreciated forum for scientific research in biomedical text mining, with an emphasis on crowdsourcing, the semantic web, knowledge integration and data linking.

The purpose of this guest editorial is to address the article contributions in this Special Issue [4] in order to use them as a basis in this area of science, whilst also encouraging communication among the various disciplines by identifying and grouping together complementary research solutions.

## 2. Challenges in the Biomedical Field

Due to the diversity of formats and text collections, to analyse biomedical data is a challenging task. Most of the online information is available nowadays as unstructured documents [5], although several initiatives have been proposed to convince users to manually label their online publications using a specialized markup language [6].

Among the hottest research topics covered by the biomedical research domain, we can mention: crowdsourcing and collaborative approaches in biomedicine in the era of big data [7,8]; creating, organizing and storing biomedical data in digital libraries [9]; extracting various information from medical texts and images [10,11]; the creation of conceptual graphs, medical ontologies and specialized semantic networks [12]; applications specially designed for medicine and biology, such as medical question answering or summarization; medical search engines [13] or deep learning techniques for bioinformatics [14]; or knowledge discovery using text mining [15].

One of the big enablers of the recent progress in the field has been an enormous growth in the amount of annotated data available. A recent publication cited 25 corpora of biomedical data that had become available within the past five years only [16,17], a previously unprecedented growth rate.

## 3. Summary

The aim of the Special Issue “Curative Power of Medical Data” of the *Data* Journal was to develop a community of researchers involved in biomedical research. This Special Issue contains four surveys, which include a wide range of topics, from author confidence in biomedical articles [18]; generating tests from medical references [19]; constructing a Gold standard biomedical corpus for the Romanian language [20]; up to the visualization of biomedical data among the Chinese elderly [21].

Mining biomedical literature to extract the science behind it, such as concepts, patterns or relations, is a very productive research area. However, extracting non-scientific information from biomedical data has recently also seen an increasing interest, with applications ranging from identifying speculative language, to retrieving papers with a specific writing style, in an attempt to cope with different reader preferences. Aside from mining knowledge, a new research direction, presented in [18], tries to identify the factors that drive readers to choose one scientific article instead of another.

Another direction proposed in the Special Issue was the automatic generation of tests for the evaluation of a student’s academic performance in the medical domain. Using medical reference texts as input and supported by a specially designed medical ontology, [19] propose a system that generates different types of test questions (multiple-choice, fill in the blanks, true/false, and match), which can have a customizable length and difficulty, and most importantly can be automatically graded.

Besides the development of tools, this Special Issue also introduces a resource, i.e., the first morphologically and terminologically annotated biomedical corpus of the Romanian language [20]. With almost 14,000 tokens distributed in three medical subdomains (cardiology, diabetes and endocrinology), the corpus contains manually validated parts of speech and named entity annotations, useful for training specific biomedical applications.

---

<sup>1</sup> The 2nd Workshop on Curative Power of MEDical DAta (MEDA) was held on 6 June 2018 at Fort Worth, Texas as a workshop at JCDL 2018.

Another research direction is studied in [21], aiming to explore, through a visual analysis, the factors which contribute to the life satisfaction of elder people, ranging from demographic, physiological, psychological and economic perspectives up to social characteristics. The findings revealed that the influence of most factors has been changing in the past ten years. The influence of bright personality has in particular been increasing since 2005, while the influence of medical accessibility has been declining. At the same time, self-rated health, the self-evaluation of the economic level, economic self-sufficiency and bright personality have proven to be the most important factors.

**Author Contributions:** Writing, review and editing D.G., D.T., K.C., and J.X.

**Funding:** This research was partially supported by a grant from the Romanian Ministry of Research and Innovation, CCCDI-UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/73PCCDI (ReTeRom), within PNCDI III and by the README project “Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users; writing styles”, contract no. 114/15.09.2017, MySMIS 2014 code 119286.

**Acknowledgments:** The authors gratefully acknowledge the administrative and technical support of the *Data* team.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Gifu, D. Malaria Detection System. In Proceedings of the International Conference on Mathematical Foundations of Informatics MFOI-2017, Chisinau, Republic of Moldova, 9–11 November 2017; Cojocaru, S., Gaidric, C., Druguş, I., Eds.; Academy of Sciences of Moldova: Chişinău, Republic of Moldova, 2017; pp. 74–78.
- De Keizer, N.F.; Abu-Hanna, A.; Zwetsloot-Schonl, J.H.M. Understanding terminological systems I: Terminology and typology. *Methods Inf. Med.* **2000**, *39*, 16–21. [PubMed]
- Cornet, R.; de Keizer, N.F.; Abu-Hanna, A. A framework for characterizing terminological systems. *Methods Inf. Med.* **2006**, *45*, 253–266. [PubMed]
- Gifu, D.; Trandabat, D.; Cohen, K.; Xia, J. Special Issue on Curative Power of Medical Data. Available online: [https://www.mdpi.com/journal/data/special\\_issues/MEDA2018](https://www.mdpi.com/journal/data/special_issues/MEDA2018) (accessed on 14 June 2019).
- Holzinger, A.; Stocker, C.; Ofner, B.; Prohaska, G.; Brabenetz, A.; Hofmann-Wellenhof, R. Combining HCI, Natural Language Processing, and Knowledge Discovery—Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data, Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 13–24.
- Guo, J.; Takada, A.; Tanaka, K.; Sato, J.; Suzuki, M.; Suzuki, T.; Nakashima, Y.; Araki, K.; Yoshihara, H. The development of MML (Medical Markup Language) version 3.0 as a medical document exchange format for HL7 messages. *J. Med. Syst.* **2004**, *28*, 523–533. [CrossRef] [PubMed]
- Khare, R.; Good, B.M.; Leaman, R.; Su, A.I.; Lu, Z. Crowdsourcing in biomedicine: Challenges and opportunities. *Brief. Bioinf.* **2015**, *17*, 23–32. [CrossRef] [PubMed]
- Shaikh, A.R.; Butte, A.J.; Schully, S.D.; Dalton, W.S.; Khoury, M.J.; Hesse, B.W. Collaborative biomedicine in the age of big data: The case of cancer. *J. Med. Internet Res.* **2014**, *16*, e101. [CrossRef] [PubMed]
- Banks, M.A.; Peay, W.J. The Future of Biomedical Digital Libraries. *Biomed. Digit. Libr.* **2006**, *3*, 5. [CrossRef]
- Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.; Granton, P.; Aerts, H.J. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446. [CrossRef] [PubMed]
- Meystre, S.M.; Savova, G.K.; Kipper-Schuler, K.C.; Hurdle, J.F. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb. Med. Inf.* **2008**, *35*, 128–144.
- Shi, F.; Chen, L.; Han, J.; Childs, P. A Data-Driven Text Mining and Semantic Network Analysis for Design Information Retrieval. *J. Mech. Des.* **2017**, *139*, 111402. [CrossRef]
- Büttcher, S.; Clarke, C.L.; Cormack, G.V. *Information Retrieval: Implementing and Evaluating Search Engines*; MIT Press: Cambridge, MA, USA, 2016.
- Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Brief. Bioinf.* **2017**, *18*, 851–869. [CrossRef] [PubMed]

15. Westergaard, D.; Stærfeldt, H.H.; Tønsberg, C.; Jensen, L.J.; Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comp. Biol.* **2018**, *14*, e1005962. [[CrossRef](#)] [[PubMed](#)]
16. Cohen, K.B.; Verspoor, K.; Fort, K.; Funk, C.; Bada, M.; Palmer, M.; Hunter, L.E. The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain. In *Handbook of Linguistic Annotation*; Springer: Dordrecht, The Netherlands, 2017; pp. 1379–1394.
17. Cohen, K.B.; Goss, F.R.; Zweigenbaum, P.; Hunter, L.E. Translational Morphosyntax: Distribution of Negation in Clinical Records and Biomedical Journal Articles. *Stud. Health Technol. Inf.* **2017**, *245*, 346–350.
18. Onofrei Plămadă, M.; Trandabăţ, D.; Gîfu, D. Towards Identifying Author Confidence in Biomedical Articles. *Data* **2019**, *4*, 18. [[CrossRef](#)]
19. Pistol, I.; Trandabăţ, D.; Răschip, M. Medi-Test: Generating Tests from Medical Reference Texts. *Data* **2018**, *3*, 70. [[CrossRef](#)]
20. Mitrofan, M.; Barbu Mititelu, V.; Mitrofan, G. Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language. *Data* **2018**, *3*, 53. [[CrossRef](#)]
21. Zhang, H.; Wang, Y.; Wu, D.; Chen, J. Evolutionary Path of Factors Influencing Life Satisfaction among Chinese Elderly: A Perspective of Data Visualization. *Data* **2018**, *3*, 35. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).