EJBI 2016 ISSN 1801 - 5603

European Journal for Biomedical Informatics

Volume 12 (2016), Issue 2

Editor

Jana Zvárová

An Official Journal of the European Federation for Medical Informatics

www.ejbi.eu

Open Data in Health and Clinical Care

Anna M. Bianchi[1], **Jana Zvárová**[2]

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

[2] Institute of Hygiene and Epidemiology, 1st Faculty of Medicine, Charles University, Prague, Czech Republic

The Volume 12, issue 2, 2016 of the European Journal of Biomedical Informatics deals with the special topic **Open Data in Health and Clinical Care** and other two topics of biomedical informatics. The gap between the demand for healthcare from an increasingly well-informed citizens and the ability of the government and healthcare organizations to meet this demand is widening all the time. In a complex environment such as the one related to health, which includes public and private sectors, services to the citizens as well as research, and addresses many different actors at many different levels, the issue of Open Data opens new challenging perspectives, but, at the same time, poses questions and problems which require to be adequately faced by the International community. From one side, new National and International regulations, are going towards the policy of making public the data collected for research purposes. On the other side, the great advances and progresses in ICT as well as the availability of wearable technologies, which allow the remote and continuous monitoring of people in different circumstances and situations, are able to generate large amounts of data. Further, clinical and administrative data collected in the public health systems, cover also different aspects related, for example, to administrative information, quality of the services, etc. All these are, with no doubt, precious sources of information available for creating further research or for improving public services or even for different purposes. The so called "data re-use" obviously relates to different aspects which include not only interoperability, privacy and security, but also strategies for preventing from misinterpretation of the actual meaning of the original data.

Thus, Bonacina, in his paper titled **Linked Open Data in Health and Clinical Care: A Review of the Literature**, goes beyond the mere concept of shared data, and also associates other fundamental resources: shared ontologies, knowledge bases, and datasets, just for mentioning a few, and presents the paradigm of Linked Open Data with the perspective of increasing the potentiality of the research community.

A completely different scenario is presented in the paper **Open Data in the Health Context: The Lombardy Region Experience**, by Barone and coauthors. Here the point of view of a Government institution for public health is faced. The need of fulfilling National and EU regulations, and the examples from other Countries with stronger traditions in this field, pushed towards the creation of services addressed to the citizens, in order to provide them the tools for a personal and more informed use of the public resources. On the other hand the authors are also well aware of the importance of making these data available for research and foresee new future uses.

The paper **Gregor Mendel's Genetic Experiments: A Statistical Analysis After 150 Years** by Kalina proposes new two-stage statistical models, which are in a better accordance with Mendel's data than a classical model, where the latter considers a fixed sample size. If Mendel realized his experiments following such two-stage algorithm, which cannot be however proven, the results would purify Mendel's legacy and remove the suspicions that he modified the results. Mendel's experiments are described from a statistical point of view and his data are shown to be close to randomly generated data from the novel models. The paper also discusses Mendel's legacy from the point of view of biostatistics.

The last paper **Students' Behavior Related to Oral Health** by Markovic evaluates the oral health behavior of dental students of the Medical Faculty in Podgorica and the Faculty of Political Science in Podgorica. The survey instrument was a questionnaire of closed type, containing questions about oral hygiene, visits to the dentist, as well as questions about nutrition and consumption of tobacco. The study showed that students take care of their oral health, but there is a need for continuous education programs on the importance and protection of oral health.

In the present issue the subject of **Open Data in Health and Clinical Care** is only partially explored. On the other hand, it is remarkable that the contributors have quite different affiliations: the former is from a Sweden Clinical Institute, the latter involves a substantially public Italian company, owned by the Government health care institutions. Thus the interest is really widespread. There are many other technical, legal, governance and use aspects which deserves to be deeply investigated, also taking into account the points of view of the different actors and users. On the other hand it's editors' hope that the present papers may rise the interest in these subjects and stimulate further contributions from the stakeholders delivery. The editors wish all interested parties an enjoyable reading and are indebted to thank all authors and reviewers for their excellent work.

Linked Open Data in Health and Clinical Care

A Review of the Literature

Stefano Bonacina[1]

[1] Health Informatics Centre, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

Abstract

Background: In the range of Semantic Web, the idea of linking and sharing the resources generated by different authors, like ontologies, knowledge bases, or datasets, is referred to "Linked data". Then, an ambitious project within the "Linked Data" paradigm is the "Linking Open Data" community project. It aims at publishing open data sets on the Web and semantically connecting data items belonging to different data sources.
Objectives: The purpose of this paper is to present a literature review on the subject of Linked Open Data in Health and Clinical Care. In fact, the availability of open data would increase evidence of the results of biomedical research, and consequently, of clinical practice.
Methods: Selection criteria have been defined and searching in PubMed/Medline and Scopus citation databases - for all years the database were available - journals papers have been retrieved. Finally, an evaluation grid has been defined for analysing the retrieved papers, to answer some defined research questions.
Results: Nine journal articles have been analysed according to the defined evaluation grid. In five out of nine papers, the main contributions are strategies and methodologies for the integration of systems, including bridging the information gap among forms for clinical research and the one for patient care. Then, in three papers the main contributions are the development of consistent triple stores according to the "Linked Data" paradigm. Finally, the last paper aims at building an open dataset for public health purposes.
Conclusions: The review was able to answer the research questions, despite the limited number of included papers.

Keywords

Linked Open Data, Semantic Web, Healthcare, Clinical Care, Ontology

Correspondence to:

Stefano Bonacina
Health Informatics Centre; Department of Learning, Informatics, Management and Ethics; Karolinska Institutet
Address: Tomtebodavägen 18a, 171 77 Stockholm, Sweden
E-mail: stefano.bonacina@ki.se

EJBI 2016; 12(2):en2-en11
received: June 30, 2016
accepted: November 22, 2016
published: December 31, 2016

1 Introduction

The term "Open Data" refers to the opportunity of using and distributing freely available data or databases generated and shared by third parties, for own business or research [1]. The concept of openness of data implies a certain number of issues to face. In fact, applications involving "Open Data" should assure interoperability, security of the transmission and storage of data, and the continuous accessibility. In addition, updates and maintenance should be accomplished in a way that prevents from malfunctioning, misunderstanding, or misinterpretation of the original meaning of data and preserves data quality [2].

In the healthcare domain, the availability of open data would allow including considerable amount of data in processing tasks, envisaging for great evidence of the results [3]. As above mentioned above, a prerequisite is the interoperability among systems. It does not mean that a unique electronic system should be used in different institutions/companies, but the data should be represented in a standardised format in order to be correctly understood. Different standards exist and are applied in the health IT domain. Without the claim to be complete, examples of standard are Health Level Seven (HL7) Clinical Document Architecture (CDA), version 2 [4], for the communication of information systems by messages; HL7 Fast Health Interoperability Resources (FHIR) [5], for boosting effective implementations compared to the previous one; Dig-

ital Imaging and Communication in Medicine (DICOM), version 3.0, for the transmission of medical images and their characteristics among digital imaging instrumentation [6, 7]. Finally, the ISO/IEEE 11073 is a family of standards for enabling the communication among devices and computer systems, in hospital settings or at patient home [8].

When data are expressed by different formats, but share the same meaning ontologies can be used for overcoming the heterogeneous representation formats and preserving the meaning. According to Gruber, an ontology is defined as an "explicit specification of a conceptualization" [9]. The main concepts of that definition are "conceptualization" and "specification". First, a "conceptualization" is the abstract representation of what we experienced and would like to represent in terms of concepts, objects, and their relationships - including their properties [10]. Then, a "specification" is for specify the conceptualization according to a language [10]. For specifying ontologies, natural language or other kinds of languages (formal) can be used [10]. The Web Ontology Language (OWL) [11], and the Resource Description Framework (RDF) Language [12] are applied for specifying ontologies on the web that need to be shared by software systems, also without human intervention (Semantic Web) [12, 13]. In other words, those languages allow delivering data and its meaning (semantics) among heterogeneous systems on the Web, also in a distributed way. It is achieved by means of representing data by triples. A triple is composed by a Subject, a Predicate, and an Object, and represented as follows, (Subject-Predicate-Object) [12]. The meaning of triple elements is similar to the one of the elements of a sentence in natural language, e.g. "A doctor visits a patient". As more triples can refers to the same entity - intended as the value of an element, e.g. "A doctor" – a graphical representation is adopted for depicting all the information about that entity. Generally, that graphical representation is a direct graph where the subject and the object are represented as nodes of the graph, while the predicate is represented by the edge from the subject to the object [12]. The RDF language is used for representing data set, ontologies, or knowledge bases by graphs composed of triples distributed on the Web. For accessing and retrieving data from that kind of information structures expressed by RDF language, the SPARQL query language has been developed by the W3C consortium [14]. The recursive acronym SPARQL stands for "SPARQL Protocol And RDF Query Language". The term "Protocol" in the acronym is for specifying that the language includes a protocol - i.e., a set of rules - for publishing the results of a SPARQL query on the Web. Then, a SPARQL Endpoint is a software service that receives queries expressed in SPARQL language, and returns the results in accordance with the rules of the protocol. A catalogue of currently available SPARQL endpoints is maintained by W3C [15].

In the range of Semantic Web, the idea of linking and sharing the resources generated by different authors, like ontologies, knowledge bases, or datasets, is referred as "Linked data" [13]. Examples of projects that can be considered in the range of "Linked Data" are the Gene Ontology [16], and the Open Biomedical Ontologies (OBO) Foundry [17]. Then, an ambitious project within the "Linked Data" paradigm is the "Linking Open Data" community project [18]. It aims at "publishing open data sets as RDF on the Web and setting RDF links between data items from different data sources" [19]. One achievement of this continuously running project is the "Linking Open Data" cloud [20]. It is a set of interconnected datasets and other data sources expressed in RDF language and accessible by SPARQL Endpoints [15].

As we mentioned above, the availability of open data would increase evidence of the results of biomedical research [3]. So, having a view of the research involving "Linked Open Data" (LOD) in the domains of healthcare or clinical care would be of great significance.

Unfortunately, a review of scientific research involving LOD in those domains is not yet available, at least according to the author's knowledge, at the time of writing. Therefore, this paper aims at presenting a review of LOD scientific research projects in Health and Clinical Care. To this end, searches in literature databases (i.e. PubMed/Medline and Scopus) have been performed and the retrieved papers analysed. In analysing the paper, the properties of interoperability [21], security of the transmission and storage (protection), and the availability (continuous accessibility) and maintenance have mainly been considered.

2 Methods

2.1 Selection criteria and literature search

The main questions guiding this review are as follows,

1) which research in Healthcare and Clinical care involves "Linked Open Data" sources?
2) How have the issues of interoperability, protection, accessibility (availability), and maintenance been faced?
3) Which are the main contributions and the envisaged directions for future research?

The following inclusion criteria were defined for including studies about "Linked Open Data" in this review:

1) "Linked Open Data" in "Healthcare" or "Clinical care" is the main subject,
2) journal articles – including the "in press" ones (historical articles, review articles, editorials, book chapters, conference papers, and grey literature were excluded),
3) articles published before the date of literature searches,
4) duplicates were included once, and finally,

5) articles written in English.

The sources considered for the online searches were Medline/PubMed and Scopus citation databases. In both PubMed and Scopus databases the "Linked Open Data" search string was searched in title, abstract, or keywords of a citation. In more detail, search strings were defined including "Linked Open Data" as a part of the title, a keyword, or a term in the abstract and combining "Healthcare" or "Clinical care" terms, respectively. The search of those terms was not limited to a specific section of a publication, e.g. "All fields" tag included for searching in PubMed database.

On April 20th, 2016, literature searches were performed in Entrez, the National Center for Biotechnology Information (NCBI) search and retrieval system (PubMed.com) and in Scopus database (Scopus.com) according to the criteria above mentioned. Table 1 presents the query strings used in both of the considered sources. Figure 1 depicts the flow chart of selection criteria. Both of them include the figures of the search results. Every article was examined by the author according to the methods described in the next section.

2.2 Classification grid for literature characteristics

Considering the three research questions above mentioned, we defined the attributes for a classification grid

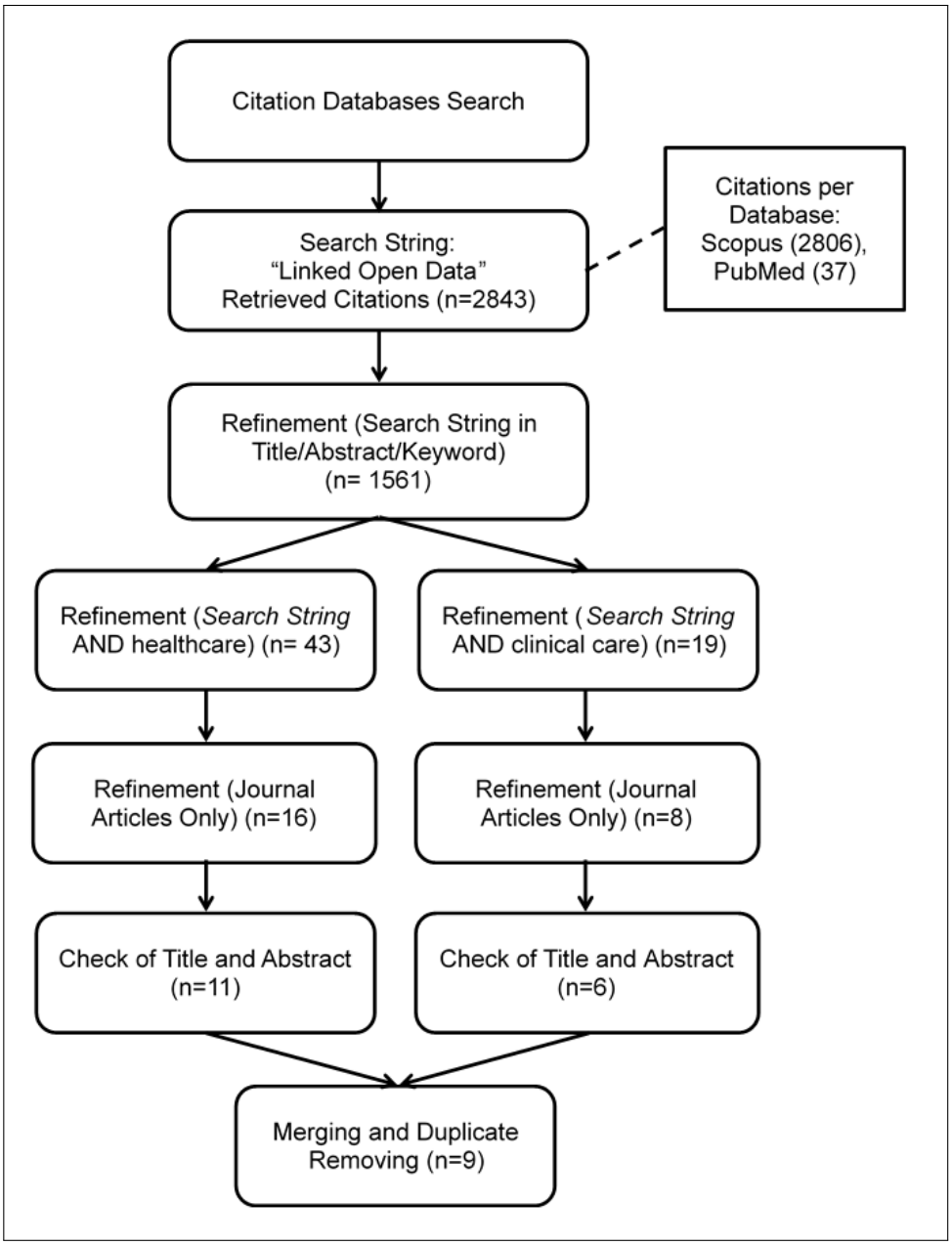

Figure 1: Literature search process according to the selection criteria.

Table 1: The search strings used and the number of identified citations.

No.	Search String	PubMed Results	Scopus Results	Sum of Results	Distinct Results
1	"Linked Open Data"[All Fields]	37	2806	2843	-
2	"Linked Open Data"[TIAB, KEYWORDS]	37	1524	1561	-
3	"Linked Open Data"[TIAB, KEYWORDS] AND healthcare	6	37	43	-
4	"Linked Open Data"[TIAB, KEYWORDS] AND healthcare (only journal articles)	4	12	16	13
5	"Linked Open Data" [TIAB, KEYWORDS] AND clinical care	1	18	19	-
6	"Linked Open Data" [TIAB, KEYWORDS] AND clinical care (only journal articles)	1	7	8	7

of the retrieved paper satisfying the inclusion criteria. For answering research question, 1) we defined the following attributes, "Article title", "Aim of the Research", and "Data Sources or Linked Open Data Involved" in the research. Generally, "Article title" and "Aim of the Research" attributes are easy to identify/extrapolate from a paper, while "Data Sources or Linked Open Data Involved" requires a scanning of the paper. That attribute refers to the data sources proposed or used in the research, including the ones available in the LOD cloud [20].

For answering research question 2), we defined the following attributes, "Interoperability / Accordance with Linked Open Data principles", "Protection", "Accessibility (availability), and maintenance". A brief explanation of their meaning is as follows. Interoperability is the ability of system to interact and understand each other, i.e., exchanging data, without any "special effort" [21]. In the LOD field, some "linked data" principles have been proposed by Berners-Lee [13, 22]. Those principles classify data according to a five-star scale, as follows [22]: 1) Data is available in the web according to the Open licence (no matter about the format); 2) as 1), plus data is expressed in machine-readable format (e.g. spreadsheets, tables as images); 3) as 2), plus data is expressed in non-proprietary format (e.g. Comma Separated Values – CSV format); 4) as 3), plus data is expressed according to W3C open standards (e.g., RDF and SPARQL) and they can be accessed by those standards; 5) as 4), plus "Linked RDF", "Link your data to other people's data to provide context" [22]. Then, "protection" refers to the tools or strategies adopted for protecting data, including the preservation of anonymity for data obtained from patients. Further, "Accessibility (availability), and maintenance" refers to the strategies adopt for ensuring a longitudinal service. In case of "5 stars data", it includes methods for updating data to maintain "Linked RDF" when data from others has been changed.

Finally, for answering research question 3), we defined the following attributes, "Main Contributions", and "Directions for future research". The "Main Contributions" attributes presents a summary of the findings considering the point of views of the healthcare domain and LOD domain; "Directions for Future Research" briefly presents next steps for contributing to the LOD domain. Furthermore, the "Year of Publication" has been considered for sorting the articles included in the review, and the "Reference Number" gives the citation number in the References section.

Summarising, the Classification Grid for literature characteristics is composed by the following attributes, "Reference Number", "Year of Publication", "Article Title", "Aim of the Research", "Data Sources or Linked Open Data Involved", "Interoperability / Accordance with Linked Open Data principles", "Protection", "Accessibility (availability), and Maintenance", "Main Contributions", and "Directions for Future Research".

3 Results

As reported in Table 1, and in Figure 1, the literature search identified 37 citations in PubMed and 2806 citations in Scopus applying ¡"Linked Open Data"[All Fields]¿ as a search string. Specifying the search strings according to the aim of the review, and the inclusion criteria (Figure 1), 9 papers have been considered and examined [23–31].

The results of the paper examination are presented in the Tables 2, 3, and 4, as the attributes for answering research questions 1, 2, and 3, were grouped as mentioned above. In Table 2, the attributes for answering research question 1 are presented for each paper. In 4 out of 9 papers, a framework – or system – supporting LOD datasets is developed locally for specific purposes (papers 2, 3, 5, and 7) [24, 25, 27, 29], see Table 2. In papers, 6, 8, and 9 the aim is the development of ontologies in RDF format [28, 30, 31]. Finally, in the paper 1 [23], a database has been developed with open data access, without Semantic Web technologies. As for the data sources involved, papers 2, 3, 5, and 7 [24, 25, 27, 29], include data collected locally and LOD sources to use for integrating or enriching those data in RDF format. In papers, 6, 8, 9, some LOD ontologies have been used as a base for the developed ones [28, 30, 31]. In paper 1, LOD sources have not been used (Table 2). Finally, the publication year for 3 out of 9 papers (papers 1, 2, and 3) [23–25] is 2013, for 3 out of 9 papers (papers 4, 5, and 6) [26–28] is 2014, for 1

out of 9 papers (paper 7) [29] is 2015, and for 2 out of 9 papers (paper 8, and 9) [30, 31] is 2016, Table 2.

In Table 3, the attributes for answering research question 2 are presented for each paper. In 4 out of 9 papers, five-star-linked open data requirements are met (papers 2, 3, 6, and 8) [24, 25, 28, 30] and a link available to the linked data datasets/system developed and presented in the paper have been found on the web, (Table 3, "Accessibility (availability), and maintenance" attribute). In three cases, papers 5, 7, and 9 [27, 29, 31], four-star-linked open data requirements would meet if the linked data datasets/system were published on the web, (Table 3, "Accessibility (availability), and maintenance" attribute). In one case, for paper 1 [23], three-star-linked open data requirements are met, as the developed Domesday dataset is available as Excel file and CSV format, (Table 3, "Accessibility (availability), and maintenance" attribute). Finally, paper 4 [26], does not meet linked open data requirements as a tool with just exemplary data is available (Table 3, "Accessibility (availability), and maintenance" attribute). As for patient data protection, it appears that it was properly addressed in 4 out of 9 papers (papers 1, 3, 7, and 8) [23, 25, 29, 30], while it appears not be faced in 5 out of 9 papers (papers 2,4,5,6, and 9) [24, 26, 27, 31].

In Table 4, the attributes for answering research question 3 are presented for each paper. In 5 out of 9 papers (papers 2,3,4,5, and 7) [24–27, 29] the main contributions are strategies and methodologies for the integration of systems, including bridging the information gap among forms for clinical research and the one for patient care (Table 4, "Main Contributions" attribute). Then, in 3 out of 9 papers (papers 6,8, and 9) [28, 30, 31] the main contributions are the development of consistent triple stores (Table 4, "Main Contributions" attribute). As for "Directions of Future Research" attribute (Table 4), performance improvement or increasing dataset size are the most mentioned.

4 Discussion and Conclusions

The presented review gives a snapshot of the current state of the research in the field of Linked Open Data for Healthcare and Clinical care. Tables 2, 3, and 4, describe the research performed in the field, how the issues of interoperability, protection, accessibility (availability), and maintenance have been faced, and main contributions and envisaged directions for future research, respectively.

According to the author's knowledge, at the time of writing, previous reviews on the subject of Linked Open Data in Health and Clinical Care have not been published. The number of considered papers – nine - could seem really limited; however, a rigorous and repeatable methodology has been followed in the literature search, as mentioned in the Methods section. Considering the publication year of the papers included in the review, two were published in the first third of the year, so more papers could be published on the subject, by the end of the year. Another reason for the limited number of papers respecting the inclusion criteria is the complexity of the Web Semantic subject and its limited spread in the clinical practice domain [18]. In fact, Web Semantic is still a research subject with main focus on ontologies [9, 10, 16, 17].

The results of this review revealed the main significant applications in the field developed so far [23–31]. In addition, the papers analysed have been evaluated according to the 5 star-based scales for linked open data requirements [22]. Then, the amount of papers analysed allowed answering the three research questions.

Those results can contribute to start new research initiatives in the domain of Linked Open Data for Healthcare and Clinical care. One direction is to initiate research in sub-domains that is not included in those mentioned in Table 2. Another direction is to improve current initiatives for fulfilling Linked Open Data principles comprehensively (see the Introduction), while protection, accessibility, and availability of data are maintained, as presented in Table 3. A third direction is to start research in the "Directions for Future Research", as described in Table 4.

The methodology applied in the presented review can help researchers in other fields in performing literature reviews. First, research questions have been defined. Then, literature databases have been selected according to the research field (LOD in Healthcare and Clinical care, in this literature review), search phrases have been defined, and searches executed. Then, a classification grid has been defined for being able to answer the research questions. The described process can be generalized and applied for other research fields.

Finally, this review presents some limitations. The authors decided to include only two citation databases (PubMed/Medline and Scopus), and including in the study only papers published in journals. However, considering conference proceedings would have led to the inclusion of the same project/system more than once, as conference proceedings are for presenting partial results of ongoing research.

In five years, the same literature review could be performed again and the results compared for assessing the spread of LOD in Health and Clinical Care.

Acknowledgements

The research here presented was supported by the Health Informatics Centre, Karolinska Institutet.

References

[1] The Open Definition [Online]. Available from: `http://opendefinition.org` Accessed: 15 June, 2016.

[2] Papavasileiou V, Flouris G, Fundulaki I, Kotzinos D, Christophides V. High-level change detection in RDF(S) KBs. ACM Trans Database Syst. 2013;38(1): Article 1, 42 pages. DOI=http://dx.doi.org/10.1145/2445583.2445584.

[3] Giglia E. Open access in the biomedical field: a unique opportunity for researchers (and research itself). Eura Medicophys. 2007 Jun;43(2):203-13.

[4] Raths D. Trend: standards development. Catching FHIR. A new HL7 draft standard may boost web services development in healthcare. Healthc Inform. 2014 Mar;31(2):13, 16.

[5] Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, Shabo Shvo A. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006 Jan-Feb;13(1):30-9.

[6] Kohn D. RSNA (Radiological Society of North America) participants embrace new imaging standard. DICOM 3.0 promises easy healthcare imaging communications. Health Manag Technol. 1994 Mar;15(4):26-9.

[7] Bidgood WD Jr, Horii SC. Introduction to the ACR-NEMA DICOM standard. Radiographics. 1992 Mar;12(2):345-55.

[8] International Organisation for Standardisation & The Institute of Electrical and Electronics Engineers. ISO/ IEEE 11073 family of standards. 2004. Available from: http://standards.ieee.org/ Accessed: 15 June, 2016.

[9] Gruber TR. A Translation Approach to Portable Ontologies. Knowledge Acquisition. 1993;5(2):199–220.

[10] Guarino N, Oberle D, Staab S. What Is an Ontology? In: Staab S, Studer R, editors. Handbook on Ontologies, 2nd Ed. Springer-Verlag: Heidelberg; 2009. p. 1-17.

[11] W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview (Second Edition). 11 December 2012. W3C Recommendation. URL: http://www.w3.org/TR/owl2-overview/

[12] Klyne G, Carroll JJ, editors. Resource Description Framework (RDF): Concepts and Abstract Syntax. [Online, 2003]. W3C Proposed Recommendation 15 December 2003. Available at http://www.w3.org/TR/rdf-concepts/

[13] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American. 2001;284;34-43.

[14] Prud'hommeaux E, Seaborne A, editors. SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. [Online, 2008]. Available from: http://www.w3.org/TR/rdf-sparql-query/ Last Access: 15 June 2016.

[15] W3C Wiki. SPARQL Endpoints. [Online, Updated 2016]. Available from: https://www.w3.org/wiki/SparqlEndpoints Last Access: 15 June 2016.

[16] Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. AMIA Annu Symp Proc. 2003:609-13.

[17] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5.

[18] Bizer C, Heath T, Berners-Lee T. Linked data—the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). 2009;5(3):1–22.

[19] W3C SWEO Linking Open Data community project. [Online, 2007; Update, 2016] Available from: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. Last Access: 15 June 2016.

[20] Schmachtenberg M, Bizer C, Jentzsch A, Cyganiak R. Linking open data cloud diagram 2014. [Online, 2010. Updated 2014]. Available from: http://lod-cloud.net/ Last Access: 15 June 2016.

[21] IEEE. Interoperability. Standards Glossary [Online]. Available from: https://www.ieee.org/education_careers/education/standards/standards_glossary.html Last Access: 15 June 2016.

[22] Berners-Lee T. Linked data. [Online, 2006. Updated, 2009] Available from: https://www.w3.org/DesignIssues/LinkedData.html. Last access: 15 June 2016.

[23] Reddington J. The Domesday dataset: linked open data in disability studies. J Intellect Disabil. 2013 Jun;17(2):107-21.

[24] Sinaci AA, Laleci Erturkmen GB. A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. J Biomed Inform. 2013 Oct;46(5):784-94

[25] da Silva KR, Costa R, Crevelari ES, Lacerda MS, de Moraes Albertini CM, Filho MM, Santana JE, Vissoci JR, Pietrobon R, Barros JV. Glocal clinical registries: pacemaker registry design and implementation for global and local integration–methodology and case study. PLoS One. 2013 Jul 25;8(7):e71090.

[26] Samadian S, McManus B, Wilkinson M. Automatic detection and resolution of measurement-unit conflicts in aggregated data. BMC Med Genomics. 2014;7 Suppl 1:S12.

[27] Tilahun B, Kauppinen T, Keßler C, Fritz F. Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. JMIR Med Inform. 2014 Oct 25;2(2):e31.

[28] Xu B, Xu L, Cai H, Jiang L, Luo Y, Gu Y. The design of an m-Health monitoring system based on a cloud computing platform. Enterprise Information Systems. 2015; DOI: 10.1080/17517575.2015.1053416

[29] Saleem M, Padmanabhuni SS, Ngomo AC, Iqbal A, Almeida JS, Decker S, Deus HF. TopFed: TCGA tailored federated query processing and linking to LOD. J Biomed Semantics. 2014 Dec 3;5:47

[30] Bamparopoulos G, Konstantinidis E, Bratsas C, Bamidis PD. Towards exergaming commons: composing the exergame ontology for publishing open game data. J Biomed Semantics. 2016 Feb 9;7:4.

[31] Kawazoe Y, Imai T, Ohe K. A Querying Method over RDF-ized Health Level Seven v2.5 Messages Using Life Science Knowledge Resources. JMIR Med Inform. 2016 Apr 5;4(2):e12

Table 2: The filled in classification grid for answering Research Question 1.

Article No. / Reference No.	Year of Publication	Article Title	Aim of the Research	Data Sources or Linked Open Data Involved
1 / [23]	2013	The Domesday dataset: Linked open data in disability studies	To develop a database (The Domesday dataset) about the Augmentative and Alternative Communication (AAC) devices in use in the United Kingdom.	Data about provision of AAC devices was collected from different types of public body, mainly the National Health System (NHS) Trusts, and Local Education Authority (LEA), between 2006 and 2011.
2 / [24]	2013	A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains	To develop a framework (a so called, federated metadata registry -MDR) for semantically linking the forms (Common Data Elements - CDEs) created for clinical research data with the ones used for patient care in clinical practice.	The MDR standard ISO/IEC 11179 has been translated according to the Linked Open Data principles and a triple store made available, as SPARQL Endpoint. The framework connects a number of CDEs (CDISC SHARE, HITSP, and OMOP), which are maintained by different MDRs, applying SKOS ontology. Simple Knowledge Organization System (SKOS) is for representing terminologies, thesauri and classifications according to Semantic Web technology (i.e. triple stores). Bioportal, which is an ontology repository and terminology server, offers common medical terminologies as triple stores.
3 / [25]	2013	Glocal Clinical Registries: Pacemaker Registry Design and Implementation for Global and Local Integration – Methodology and Case Study	To develop a device registry framework (applied to a pacemaker registry) for linking clinical research data with patient care data in clinical practice. Specifically, a database for reporting on pacemaker long-term outcomes has been developed.	Data is from the randomized trial ("Safety and the Effects of Isolated Left Ventricular Pacing in Patients With Bradyarrhythmias," ClinicalTrials.gov study ID NCT01717469). Data standards from ACC/AHA, CDISC, NCDR and Linked Clinical Trials (Linked CT) triple store, available in the LOD cloud.
4 / [26]	2014	Automatic detection and resolution of measurement-unit conflicts in aggregated data	To solve the problems of integration and harmonization among units of measurement, by a Semantic Web Service-based approach.	The clinical records of a patient cohort (536 unique patients) collected between 1986 and 1989 from a referral hospital in Nebraska, USA. GALEN compositional ontology, Semantic Automated Discovery and Integration (SADI) Semantic Web Service framework, Semantic Health and Research Environment (SHARE) mediator system, the Semantic Science Integrated Ontology (SIO).
5 / [27]	2014	Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation	To develop a system for representing health information, querying, and visualizing them - the so called Linked Open Health Data (LOHD) - by integrating public health data coming from different sources, applying the Semantic Web technologies.	Data from United Nations program for HIV/AIDS (UNAIDS), and the Linked Open datasets DBpedia, Bio2RDF, and LinkedCT.
6 / [28]	2014	TopFed: TCGA tailored federated query processing and linking to LOD	To translate The Cancer Genome Atlas (TCGA) in a triple store (i.e. a Linked Data version of the atlas) and to develop a query engine, TopFed, offering SPARQL endpoints for improving remote query processing and virtual data integration.	TCGA, by the National Cancer Institute, makes available the characterization and sequencing of more than 30 cancer types, from 9 thousand patients at the molecular level.
7 / [29]	2015	The design of an m-Health monitoring system based on a cloud computing platform	To design and develop a Software as a Service (SaaS) cloud system for facilitating collection, storage, and processing of personal healthcare data for monitoring applications.	Diseasome, Depedia, and Drugbank triple stores for linking to disease definitions and drug descriptions. Dbpedia for disease descriptions. Personal data from patients in the antimicrobial resistance case study.

Article No. / Reference No.	Year of Publication	Article Title	Aim of the Research	Data Sources or Linked Open Data Involved
8 / [30]	2016	Towards exergaming commons: composing the exergame ontology for publishing open game data	To develop a model for the semantic representation of exergames (i.e. games for guiding in physical exercises) using OWL, including an exergame ontology that embodies game session concepts. A SPARQL endpoint is offered, too.	Game Ontology Project (GOP) ontology, Generic Component Model (GCM) architecture framework for describing component-based system formally. Ontology of physical activity (OPA), Ontology of physical exercise (OPE), Friend of a friend (FOAF) ontology for the player description, NCI Thesaurus for the description of the muscles involved in the exercises, and "Quantity, Unit, Dimension and Type" (QUDT) collections of ontologies for representing the units of measurements of the game metrics, vCard for the information about the locations of game sessions, and SKOS for creating a terminology of common terms for exergames.
9 / [31]	2016	A Querying Method over RDF-ized Health Level Seven v2.5 Messages Using Life Science Knowledge Resources	To develop a method for converting HL7 messages into RDF data, and use linked drug databases for enriching results of SPARQL query of clinical data, including adverse drug events.	Anatomical Therapeutic Chemical Classification System (ATC), United States Pharmacopeia Classification (USP), SIDER 2, KEGG, MEDIS DRUG. Medication orders and laboratory test results of 148 thousand of unique patients from The University of Tokyo Hospital.

Table 3: The filled in classification grid for answering Research Question 2.

Article No. / Reference No.	Interoperability / Accordance with Linked Open Data principles	Protection	Accessibility (availability), and maintenance
1 / [23]	Three-star-linked open data requirements are met	Collected data are related to AAC devices; in addition the dataset has been designed and developed in a way that prevents revealing the identity of the AAC device users.	The dataset is available at: http://joereddington.com/aac-and-the-domesday-dataset/
2 / [24]	Five-star-linked open data requirements are met, for the federated MDR.	It appears protection of patient data is not faced in the study.	The federated MDR standard ISO/IEC 11179 in RDF format is available at: https://github.com/srdc/semanticMDR SALUS Common Information Model Ontology is available at: http://www.salusproject.eu/ontology/salus-cim-ontology.n3 OMOP CDM Content Entity Model Ontology is available at: http://www.salusproject.eu/ontology/omop-cdm-ontology.n3 CDA/CCD Content Entity Model Ontology is available at: http://www.salusproject.eu/ontology/hl7-cda-ontology.n3
3 / [25]	Five-star-linked open data requirements are met.	In accordance with Good Clinical Practices (GCP) and The Health Insurance Portability and Accountability Act (HIPAA) of 1996.	The Cardiac Pacemaker Clinical Trials is available at LinkedCT: http://www.linkedct.org/resource/trial/nct01717469/ However, last update was made in 2012.
4 / [26]	Tools for demonstrating the BMI example presented in the paper are available. Linked open data requirements are not met, as data are not available.	It appears protection of patient data is not faced in the study.	The BMI example is available at: http://biordf.org/MeasurementUnitsDemo/
5 / [27]	At present, the system developed is not available for use. If the data were published on the web, four-star-linked open data requirements would meet.	It appears protection of patient data is not faced in the study.	It appears that the availability of the system is limited and not public.
6 / [28]	Five-star-linked open data requirements are met, as Linked TCGA database is available.	The Linked TCGA does not include patient data, so appears that protection issues are not faced in the study.	The Linked TCGA database is available at: http://tcga.deri.ie/ The TopFed's utilities and tools are available at: http://goo.gl/rtwm6q
7 / [29]	At present, the Cloud-MHMS is applied as a case scenario for allowing general practitioners (GPs) in China to access patient data and refer them to the appropriate hospital. If the data were published on the web, four-star-linked open data requirements would meet.	Multiple tenant access control is implemented to allow data isolation and data sharing.	It appears that the availability of Cloud-MHMS is limited to GPs in China.
8 / [30]	Five-star-linked open data requirements are met, as data from game sessions have been published on the web.	Data about game sessions was published carefully obscuring any information that may reveal identity.	Exergame ontology is available at http://purl.org/net/exergame/ns#.
9 / [31]	At present, the developed methods are applied only at The University of Tokyo Hospital. If the data were published on the web, four-star-linked open data requirements would meet.	It appears protection of patient data is not faced in the study.	It appears that availability of data is for clinicians and medical staff at The University of Tokyo Hospital.

Table 4: The filled in classification grid for answering Research Question 3.

Article No. / Reference No.	Main Contributions	Directions for Future Research
1 / [23]	The availability of a dataset including the provision data of AAC devices in the UK for research, policy definition, and public health purposes.	Continuing update of the dataset including data from public bodies (not included in the initial development). The assessment of the impact of AAC apps available on tablets.
2 / [24]	The developed federated semantic metadata registry. The definition of CDEs in machine-readable format.	Locally developed CDEs could be linked with CDEs developed by standardisation bodies.
3 / [25]	The developed Cardiac Pacemaker Clinical Trials based on LOD principles.	Implementing the integration with a platform for adverse events monitoring, defining protocols for data enrichment through natural language processing (NLP) methodologies.
4 / [26]	The definition of a Semantic Web-based solution for encoding units of measurement in a machine readable way and harmonising them in (clinical) databases.	Improving performance of the solution when large datasets are examined, and including more complex pattern of units of measurements, and temporal units, as well.
5 / [27]	The developed LOD-based health information representation, querying, and visualization system which uses Linked Data tools. The evaluation test of the query system – based on SPARQL – showed friendly interfaces should be introduced for users unfamiliar with Semantic Web technologies.	Improving the query engine and its interface for users unfamiliar with Semantic Web technologies.
6 / [28]	The Linked TCGA RDF dataset (20.4 billion triples), the TopFed query engine, and the SPARQL endpoints. Evaluation of the TopFed performances for 10 different SPARQL queries.	Applying the system in clinical practice as a decision support tool for suggesting drugs treatment for cancer patients.
7 / [29]	A cloud computing structure based on multiple layers for collecting data from patient monitoring devices, transfer them to the storage layer, and make them available to healthcare professionals. The cloud solution allows to memorize data one while GPs and clinicians can access them avoiding data duplication and cost for multiple collecting/storing.	Linked data model will be used for data analysis in Hospital Information Systems.
8 / [30]	The Exergame ontology enriches semantically data from exergames and allows semantic processing.	Enriching the ontology increasing the number of sessions and the number of exergames. Performance and scalability evaluation.
9 / [31]	The method for converting HL7 messages into RDF; a potential large-scale data federation for retrieving clinical information enriched by drugs information. About 650 million RDF triples for medication orders and 790 million RDF triples for laboratory test results were produced.	The application of the proposed methods in other hospitals that are using HL7 standard for exchanging messages.

Open Data in the Health Context

The Lombardy Region Experience

Luca Augello[2], Antonio Barone[2], Daniele Crespi[2], Michele Ercolanoni[2], Ferdinando Ferrari[1], Matteo Giacomo Jori[3], Luca Merlino[1], Simone Paolucci[2], Francesca Romana Rossi[2]

[1] Regione Lombardia, Italy

[2] Lombardia Informatica S.p.A., Milano, Italy

[3] Università degli Studi di Milano, Italy

Abstract

The scope of this paper is to describe, within the complex national and regional legal context with respect to Open Data, the methodologies and opportunities for the use of such data in the health context in the Lombardy Region.

Keywords

Administrative data, Healthcare quality, Data re-use, Data protection

Correspondence to:

Antonio Barone
Lombardia Informatica S.p.A.
Address: Via Taramellii 26, 20124 Milano, Italy
E-mail: antonio.barone@lispa.it

EJBI 2016; 12(2):en12-en19
received: July 19, 2016
accepted: December 16, 2016
published: December 31, 2016

1 Introduction

In the digital age, data and information form a key resource for any social or commercial activity. A simple query for the nearest hospital or late-hour pharmacy requires access to data held and made available centrally, whose usage may allow the generation of innovative services of added-value.

Furthermore the availability of Open Data transforms the citizen from a passive subject into a person who can interact with foreknowledge and so be more greatly involved in the decisions that concern him/her. This represents a value that goes beyond mere transparency, giving the citizen the possibility of bringing his own contribution.

In a "democratic" system, the citizen should be able to know and share with others the policies and activities of the relative public administration. Transparency is not merely a question of access to data, but the possibility of sharing and re-using it.

2 Legal background

The origin of the concept of Open Data, and in particular, Open Government Data (Open Data in the public sector), that is making public sector data accessible and usable for whatever purpose, is usually associated with the legal system of North America. In fact, the first processes of public sector data availability towards citizens were developed in 2009 in the United States. The US experience first influenced the UK government, and then, in a short time, the public administrations of other important countries, like Canada and New Zealand. The readiness of common law legal systems to appreciate the potential of open government data is directly derived from the relationship between citizens and the public administration in Anglo-Saxon countries, traditionally based on the generalized access to information ("any person has a right, enforceable in court, to obtain access to federal agency records", as stated by the United States Freedom of Information Act of 1966 and in Directive 2003/98/EC on the re-use of public sector information, otherwise known as the PSI Directive. PSI Directive is an EU directive that encourages EU member states to make as much public sector information available for re-use as possible. For a more profound analysis of the directive and the principles on which it is based, see also [1–3].

Legal systems of continental Europe, on the other hand, traditionally consider access to public administration data as an exceptional event, justified only by specific circumstances, like for instance the non-observation of a citizen's right, and the consequent need to access the documents in question. It is not therefore surprising that the first approach to the subject of usage of public sector information, proposed by the European Community directive 2003/98/CE, was undeniably of great importance,

but demonstrated a certain timidity in the definition of the objectives and instruments to enable such usage. Even though the directive was formulated with the specific intent of encouraging the availability of public data, it did not contain any obligation towards the member states to implement such availability, and let them require a specific request and fee payment for information access. Thus section 1, paragraph 2 of the Italian implementation of the European directive, Decree 36/2006, defined that public administrations could choose "to authorize the re-use of documents containing public data collected, produced and reproduced for institutional purposes". This principle was partially balanced by section 1, paragraph 4 of the decree, according to which the goal of making information re-usable should be pursued according to methodologies that would guarantee the conditions of equality, adequacy and non-discrimination.

A rather more important step towards the implementation of a real open data system was made with the next directive, 2013/37/EU (which modified the aforementioned 2003/98/EC). This directive, while still allowing nations to locally limit access to some information, and guaranteeing in any case the laws of personal data protection, and intellectual and industrial property, defined an obligation towards the member states to consent re-use of public sector information. In conformance with the new policy defined by the European directive, Italian decree 33/2013, aka "the transparency decree", introduced into the Italian legal system the obligation towards public administrations to consent the re-use of information that must be made freely accessible according to the current transparency laws, in order to guarantee a correct government functioning. The decree, recognising the citizen as having specific rights with respect to transparency, defined a "right to access administrative data" that includes the possibility of using such data for purposes over and above those for which the data was originally collected.

Decree 33/2013 is not the only source of regulations that deal with open data access. The Italian Digital Administration Regulations (decree 82 of 7 March 2005), for example, defines important criteria for the correct definition of the data to be made available, and the instruments, both technical (data format) and legal (licensing), to use and transport the information.

The sharing of open data is considered to be part of a virtuous cycle, able to add value to the information assets of both private companies and public bodies, both in terms of research and development, and under the profile of market opportunity and the strengthening of new synergies, made possible by a more dynamic usage of databases. Therefore, in the context of open government data, with the objective of transparency, made possible by a greater monitoring by users of the actions of public administrations, advantages may be obtained that derive from the possibility of citizens to participate in the management of public resources, through information access and open data re-use.

However, the obligations of transparency and accessibility cannot be considered indiscriminate. The noble purposes of the doctrine of openness must necessarily be conditioned by other values guaranteed as constitutional rights, and in particular by the right to privacy and protection of personal data.

3 Open data and the protection of personal data in the health context

One of the most frequently recurring themes related to the implementation of open data is the possible conflict of such processes of data management and sharing with data protection laws. Without doubt, at least theoretically, the availability and re-use of information related to citizens could be in contrast with the rigorous limitations, defined by Decree 196/03 (Italian Personal Data Protection Regulations) and European Regulation EU 2016/679 (General data protection regulations), on the treatment of personal data. The above considerations are therefore particularly relevant if applied in the health sector, where databases generally contain sensitive data, in that it may reveal information relative to an individual's state of health.

The importance of defining a correct compromise between openness of information assets and the protection of personal data is evident, both in the policies of the European Union and those of the European data protection authorities. At National level, the personal data privacy authority drew up: "Linee guida in materia di trattamento di dati personali contenuti anche in atti e documenti amministrativi effettuato da soggetti pubblici per finalità di pubblicazione e diffusione sul web" of 2 March 2011 (in G.U. 19 marzo 2011, n. 64; doc. web. n. 1793293), and then "Linee guida in materia di trattamento di dati personali, contenuti anche in atti e documenti amministrativi, effettuato per finalità di pubblicità e trasparenza sul web da soggetti pubblici e da altri enti obbligati" of 15 May 2014, (in G.U. 12 giugno 2014, n. 134; doc. web n. 3134436). The main principle of the current legislation is that, independently of the goals pursued (including therefore those of transparency that regulate public administrations), whenever the publication of data, information and documents involves a treatment of personal data, such goals must take into account the rights, the basic freedoms and the dignity of the subjects involved.

Both of the directives 2003/98/EC and 2013/37/EU on information re-use took into consideration the obligations of protection of personal data, but the "Transparency Decree" 33/2013, at national level, defined more clearly the application of the two directives. The general premise of section 1, paragraph 2, according to which the goals of transparency must be pursued, in all cases, guaranteeing the respect of laws concerning State secrets, official secrets, and personal data protection, finds its implementation in the section 4 of the decree, which, in

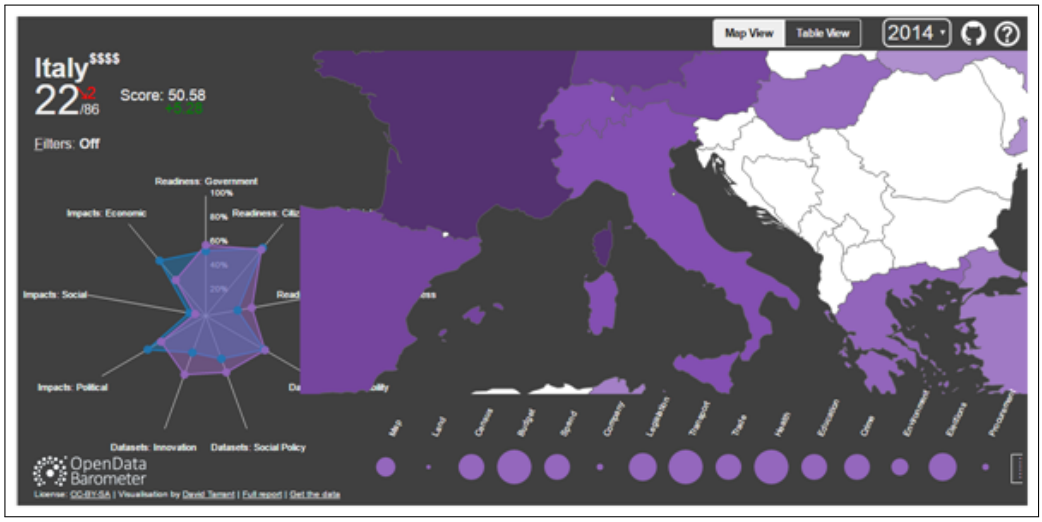

Figure 1: Open Data Barometer.

eight paragraphs defines the specifications of the "Limits to transparency". Amongst these, paragraph 1 is of particular importance. It states that personal data, when subject to publication according to the decree, must always be treated, as far as distribution, indexing and reuse are concerned, with the guarantee of the respect of regulations of the protection of personal data [4]. Paragraph 3 defines that the publishing of data, information and documents for which no such obligation of publication exists, may occur only after anonymization of any personal data present. Paragraph 4 states that, in those cases in which laws or regulations require the publication of data or documents, public bodies must render unintelligible non-pertinent personal data or unnecessary sensitive or judicial data, with respect to the specific transparency goals of publication. And again, paragraph 5, while consenting access to information relative to the actions of anyone performing a public role, defines as not exposable, unless as foreseen by law, information concerning the nature of any infirmity or personal impediment of the subject or his/her family members that have caused an absence from the workplace, or any evaluation data or information concerning the subject's job that could reveal sensitive data.

The cases disciplined by section 4 of Decree 33/2013 are not the only limitations on the publication and distribution of personal data by public bodies. Section 19, paragraph 3 of the Personal Data Protection Regulations, for example, declares the general principle according to which the communication of personal data by a public body to a private organisation or to economically autonomous public bodies, and the distribution of said data by public bodies, is allowed only when explicitly foreseen by a specific law (like for example Decree 33/2013) or regulation. Thus a public administration, before it can legally handle information assets, containing personal data, according to the criteria of open data, must first verify the existence of a law that explicitly declares that possibility.

On the other hand, the relationship between open data and the guarantee of privacy need not necessarily be one of conflict, and this is true also with reference to a delicate sector like health. Information treated as open data is often of a statistical nature, or in any case does not consent the identification of a specific citizen. The policies adopted by the Lombardy Region, and applied, for instance, to the datasets relative to the directory of pharmacies, or the analysis of hospital admissions, subdivided according to hospital, are obvious examples of how it is possible to implement processes of transparency and add value to information assets, without compromising privacy.

In this sense, data anonymization processes become particularly relevant, in that they make it possible to preserve the information value of a single datum, while removing from it any element that could allow that datum to be associated to a particular person [5]. It is useful to remember that, as defined by the Personal Data Protection Regulations, anonymous data is that which, originally or after processing, cannot be associated with an identified or identifiable subject, and to which Decree 196/03 does not apply. In the same way, point 26 of EU Regulation 2016/679 specifies that "The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable person, or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

The techniques of transformation of data into an anonymous form can be applied to so-called biosignals, whether these be signals really produced biologically, or the products of the interaction between the organism and an external agent (like, for instance, radiography, ultrasound, MRI's, electrocardiograms, ...), facilitating thus the management of such data according to the criteria of open data. The nature of such data, which carries information that is intrinsically personal, requires par-

ticular attention in the process of evaluating the status of anonymity. While the results of generic blood analysis, simply disassociated with the person's identity can be reasonably considered to be anonymous, the same cannot be said in the case of bioimaging, which, in certain circumstances, could permit the identification of the person involved [6, 7].

4 Open data in Italy. Distribution and usage in the health context

In the general context of the participation of the citizen in the events that concern him/her, the concept of empowerment, obtained through the acquisition of knowledge and the distribution of information, becomes a key element.

The Presidency of the Council of Ministers, together with the Ministry of Economic Development and the Italian Digital Agency, has presented a national plan – "Strategy for Digital Growth 2014-2020", in which the "open data" platform goes in that direction.

Strictly speaking, not all Public Administration data can be considered open data. It is necessary to verify certain conditions; that is, data is open if it is:

- Available in raw, non-aggregated format;
- Available according to the terms of a licence that permits its re-use, even for commercial purposes (Italian Open Data License v2.0 or the international Creative Commons 4.0);
- Accessible through information and communication technologies in open format (i.e. a documented and technology-neutral public format);
- Available at no cost, or at most at low cost to take into account reproduction and distribution.

The Open Data Barometer of the World Wide Web Foundation provides statistics on the level of adoption of open data policies throughout the world. Italy comes out both positively, considering that it is positioned in the high part of the chart (22nd of 77), and negatively, considering the distance from the leading countries. Figure 1 shows a screenshot of the Open Data Barometer statistics.

The portal of the national Public Administration, `http://www.dati.gov.it/`, managed by AGiD, contains 10,348 datasets coming from 76 administrations. Of these only 256 are health-related.

There are just 30 datasets of the Health Ministry portal `http://www.dati.salute.gov.it`. Table 1 shows the most popular downloads from this site.

Table 1: Downloads from `http://www.dati.salute.gov.it/`.

Dataset	Downloads
Pharmacies	54,301
Parapharmacies	26,132
Medical devices	47,172
Pharmaceutical Distributors	21,473

The potential of open data in the Health context is however in the forefront. During the event "HACK4DIGITALGOV: PA in your hands" of 2015, of the 3 ideas receiving awards, the first 2 regarded health themes, and used open data of a sanitary nature.

In particular, the first classified was the app "First Aid 2.0", which had the objective of communicating the overcrowding status of Emergency Rooms according to White, Green and Yellow codes.

In second place was "Health Advisor", with the scope of letting users express evaluations of the national health service according to their own experience.

5 Open data in the Lombardy region. Health data available

The Lombardy Open Data portal (`www.dati.lombardia.it`) forms part of the process defined by the Lombardy Digital Agenda (approved by the Regional Government Resolution IX/2585 of 30 November 2011), and in particular is one of its priority initiatives: the valorisation of the public information assets.

The objective is to make available to everyone the Region's information assets, such that large amounts of information of public interest (e.g. health data, demographic data, maps, economic, environment and climate data, ...) can be used by anyone to create new services, perform studies, create applications.

This is the added value: starting from the data that has been made available, it is possible to create innovative services for the Region, favouring not only transparency, but also participation and the collaboration between institutions and the private sector.

Four years later, with its 1500 published datasets relative to 20 different contexts, the Lombardy Open Data portal, based on the platform Socrata, is one of the most active in Italy, and is characterised by the quality of the data, both in terms of completeness and speed of update.

The usage of the Open Data is increasing constantly, as seen in the Table 2.

Table 2: Increase in page visits.

2014	2015
1,566,000 pages visited	3,341,000 pages visited (+113%)
2,814,000 lines accessed	5,336,000 lines accessed (+89%)
49,431 downloads	97,777 downloads (+102%)

The graph in Figure 2 shows visualisations and downloads (data through May 2016).

The Region has made the platform available also to other local authorities of Lombardy. It currently collects data of around 15 bodies, and 3 of these (the city of Bergamo, the city of Monza, and the Province of Monza-Brianza) have activated dedicated micro-sites.

The platform offers several advanced functions: citizens, developers and researchers can consult data, build

maps and graphs and save them on the platform in order to enrich the catalogue. Furthermore, Lombardy Region has promoted various initiatives with the scope of encouraging the use of the data, stimulating digital creativity, in particular for young people, with the participation of schools and universities.

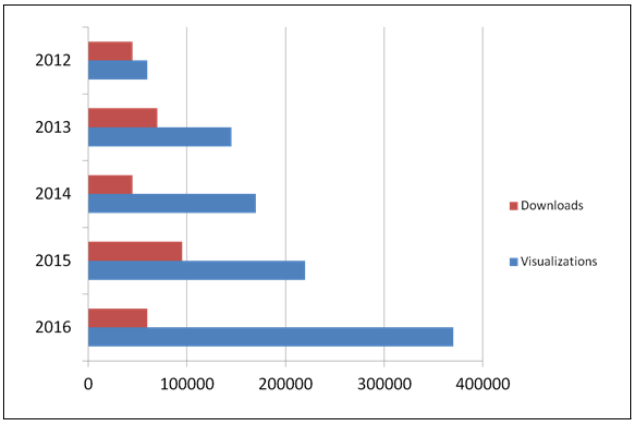

Figure 2: Visualizations (blue) and downloads (red) related to the Open Data portal in the last five years.

In the health context, 25 datasets are available (making a total of 61 taking into account the entire Welfare context), and a similar number of different views of the data, with various filters or applied to maps, which have been created by users of the portal and shared with all visitors.

Since their first publication through to May 2016, health data has been viewed 55K times, with more than 11K downloads; the same figures for the first 5 months of 2016 being 7640 and 2900 respectively.

Table 3 shows, for the most frequent keywords, the number of datasets available.

Table 3: Number of available datasets for the most frequent keywords.

Admissions	13
DRG	8
Repetitions	7
Pharmacies	6
Performance	6
Day-hospital	6
Pharmacy	6
Hospitals	6

As well as simple directory datasets (lists of hospitals, lists of pharmacies, ...), Lombardy offers performance data using the principle indicators of health episode outcome (repeated admission and return to operating room being the most important).

In fact, these queries, along with those relative to hospital productivity data, are amongst the first places of accessed data (cumulative data updated in May 2016) as shown in Figure 3.

Using the published datasets, users have made views based on filters or applied to maps. For example, the maps of all the pharmacies of the region, shown in Figure 4, or just those of the city of Milan, shown in Figure 5, based on the dataset that contains the complete index of the pharmacies in the region.

Further, in compliance with the open philosophy, it is also possible to share and discuss the results so obtained with a simple, user-friendly interface shown in Figure 6.

6 Usefulness of data opening in the Lombardy health context

The usage of open data in the context of clinical studies is only at the beginning.

Clinical studies require a level of detail in data that cannot be found in data organised for statistical purposes, and so it is necessary to make available data with a sufficient level of granularity. To do so requires a significant effort in the processing of data to make it anonymous (for example, the application of hash functions that, while still guaranteeing the correct association between records associated with the same individual, do not permit his/her identification).

Many and complicated are the processes within the Health context. These generate a vast amount of information, which are archived, not only within hospitals, but also within databases of the national public administration. Many different types of organisation or groups may benefit from the use of health data, even though it is not possible, a priori, to foresee where this added value will be created. In spite of this situation, the Lombardy Region has started in its path towards the opening of its health and administrative information assets to external bodies associated with research organisations and university studies.

The Region, within its institutional functions of the healthcare of its citizens, needs to perform activities of evaluation and monitoring of the effectiveness of the health treatments provided, appropriateness, quality of assistance, user satisfaction, and health risk factors.

The treatment of data performed by Lombardy Region has the objective of evaluating and comparing (between groups of people or hospitals) the appropriateness, effectiveness and efficiency of the healthcare provided, also taking into account specific illnesses or health problems, exposure to risk factors, the reconstruction of diagnostic, therapeutic and care pathways, and the analysis and comparison of healthcare outcomes.

This treatment of data has particularly important goals of public interest, in terms of planning, monitoring, evaluation and appropriateness of healthcare, as foreseen by section 85, paragraph 1b of Decree 196/2003.

One example of open data usage is the portal “Dove e come mi curo” – “Where and how to get well”. This portal (see Figure 7) uses public data to help patients choose the most adequate hospital for some specific medical treatment.

Figure 3: Screenshot showing the most popular datasets.

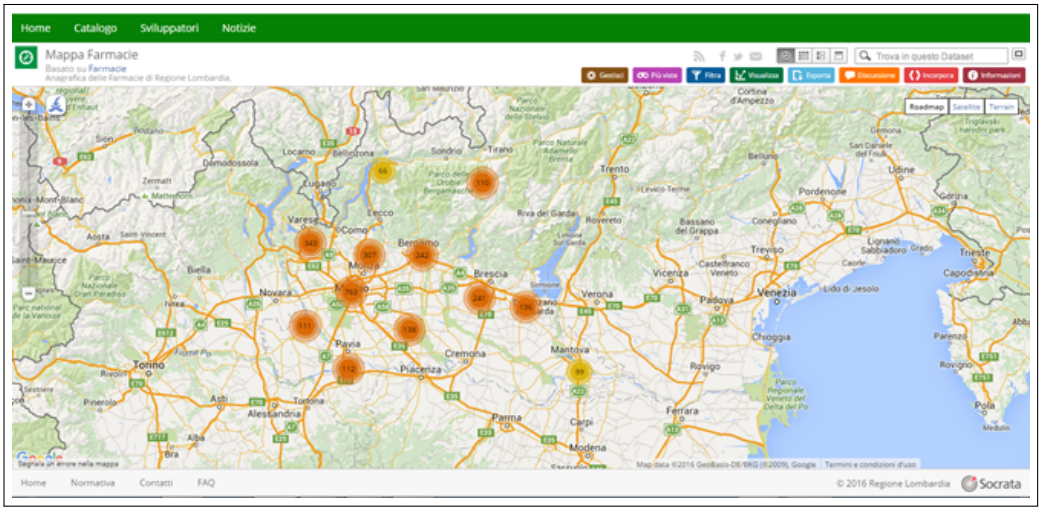

Figure 4: The map of pharmacies in the Region, grouped by Province.

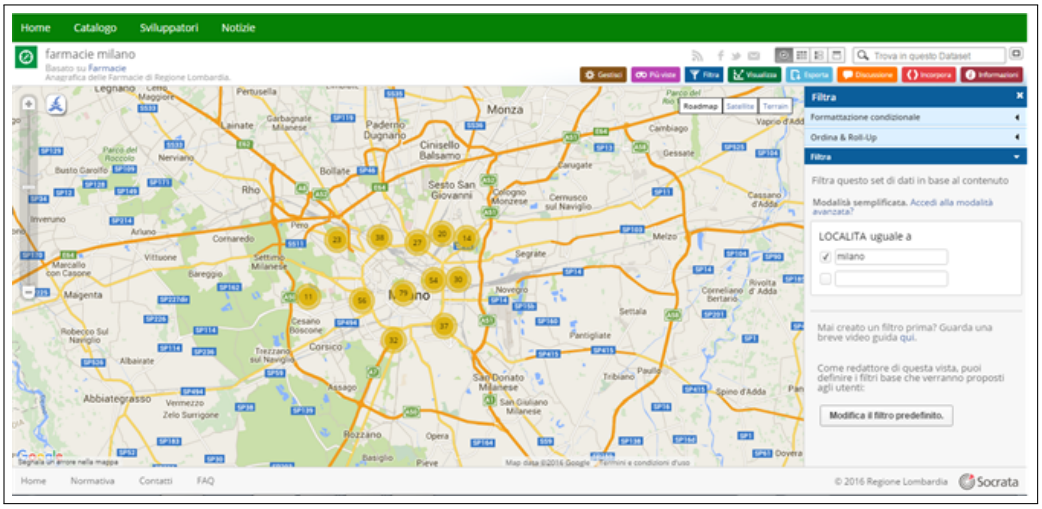

Figure 5: The map of pharmacies in the city of Milano.

Another example is the use of the georeferential coordinates of hospitals, together with Emergency Room real-time usage data, in order to allow users to locate the nearest hospital together with an estimate of its state of overcrowding. The app, SALUTILE PS, implemented in Lombardy, brings together these two datasets.

Figure 6: Interface for sharing and discussing the results obtained through the queries.

Figure 7: The portal (`www.doveecomemicuro.it`).

In order to achieve these goals, Lombardy Region foresees the participation of public and private Universities, public and private Research Hospitals, located within Lombardy and possessing specific requirements of competence in the contexts of epidemiology and research, and with high expertise in those areas[1].

These institutions, in order to collaborate with the Lombardy Region, must submit an application, which will be properly assessed and will allow their inclusion within a specific register.

Once the areas of interest related to a study to be performed have been defined, Lombardy Region chooses the body which, according to its profile in the register, is the most suitable to support it in the activities.

The involvement of these bodies is useful and of fundamental importance considering their profiles of great competence relative to the goals of evaluation of the effectiveness, appropriateness and quality of health-related activity.

The data that is made available concerns in particular:

- contagious and common diseases
- vaccinations
- early diagnosis programmes
- general practitioners
- specialist and rehabilitative outpatient care
- home care
- overseas care
- mental health
- dependencies
- hospital care
- emergency care
- residential and semi-residential care
- midwifery certificates and pregnancy outcomes
- pharmaceutical care and pharmacovigilance

[1]DGR N° X / 4893 on 07/03/2016 – "Disciplina delle collaborazioni di enti esterni con regione Lombardia nell'ambito delle attività di programmazione, gestione, controllo e valutazione dell'assistenza sanitaria, previste dall'art. 85, comma 1, lettera b) del d.lgs.196/2003 e del conseguente accesso ai dati del Data Warehouse Regionale"

- physical and sporting activities
- integrated care
- thermal care
- accident and health risks related to the workplace
- road accidents
- disabilities and handicap
- exemption from healthcare cost payment
- customer satisfaction research
- mortality data
- prosthesis

all without any elements directly identifying the patients involved (name, surname, tax code, healthcare code) such that each individual is assigned a specific unique identifier that does not consent the direct identification of the person during the treatment of data.

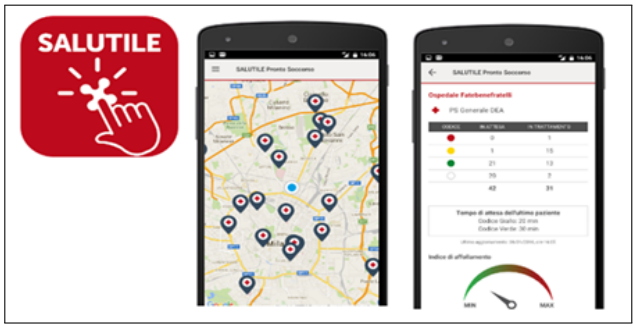

Figure 8: Emergency Room locator.

Furthermore, additional measures of data generalisation are defined and adopted in order to guarantee the non- identifiability of patients, such as using an age-range instead of the date of birth, or a province or geographical area instead of the town of birth.

The information will be used, for processing by the external bodies involved, within a specific IT platform, realised with the necessary measures of security that does not allow data to be exported from it.

The results obtained from the processing of data by the external bodies are owned exclusively by Lombardy Region, which reserves the right to consent the use of the results obtained in aggregated format (for publications, statistical comparisons with other information coming from other regions or countries) according to specific formal agreements, defined within the context of a convention stipulated before the start of the activity.

7 Conclusions

Most health-related datasets available in open data format are indexes of sanitary structures and services. As has already been said, Lombardy is one of the few regions to offer data on the evaluation of its health structures, and has in its objectives the increase in the range of datasets, which however can only happen with the necessary attention to the concerns of privacy already discussed.

The open data philosophy brings us to make ever more datasets available, and the fact of their availability stimulates the creativity of new uses in new systems and applications previously unimaginable. It must however be noted that over and above the increase in the number and quality of available datasets, it is necessary to stimulate the market that is not as yet aware of the resources that it could use; the culture of open data must be encouraged to grow within public bodies and private enterprises.

One of the more interesting areas, with recently a high growth rate, is in the integration of wearable devices (IoT – Internet of Things), equipped with a vast range of sensors able to measure the elementary physical parameters of an individual. Experiments in this field are already underway and regard, for example, remote monitoring of vital signs for patients in homecare.

References

[1] G. DE MINICO, Gli open data: una politica costituzionalmente necessaria?, in Forum di quaderni costituzionali rassegna, 12.6.2014.

[2] B. PONTI, La trasparenza amministrativa dopo il d.lgs.14 marzo 2013, Maggioli editore, Santarcangelo di Romagna, 2013, p. 115.

[3] F. MERLONI, La trasparenza amministrativa, Giuffrè, Milano, 2008, p. 364.

[4] C. ROMANO, Open data e riutilizzo nel decreto trasparenza: propulsore per la democrazia e lo sviluppo o sfida ulteriore per i diritti fondamentali?, in L. CALIFANO, C. COLAPIETRO (a cura di), Le nuove frontiere della trasparenza nella dimensione costituzionale, Editoriale Scientifica, Napoli, 2014, pp. 263-277.

[5] ARTICLE 29 DATA PROTECTION WORKING PARTY, Opinion 05/2014 on Anonymization Techniques, 0829/14/EN WP216, adopted 10 April 2014.

[6] G. PALIWAL, A.W. KIWELEKAR, A comparison of mobile patient monitoring systems, in G. HUANG, X. LIU, J. HE, F. KLAWONN, G. YAO, Health Information Science, Second International Conference, Springer, New York, 2013, 205-206.

[7] W.J. SONG, S.H. SON, M. CHOI; M. KANG, Privacy and security control architecture for ubiquitous RFID healthcare system in wireless sensor networks, in Digest of Technical Papers International Conference on Consumer Electronics, 2006, 239-240.

Gregor Mendel's Genetic Experiments

A Statistical Analysis after 150 Years

Jan Kalina[1]

[1] Institute of Computer Science CAS, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic

Abstract

Gregor Mendel is generally acknowledged not only as the founder of genetics but also as the author of the first mathematical result in biology. Although his education had been questioned for a long time, he was profoundly educated in botany as well as physics and in those parts of mathematics (combinatorics, probability theory) applied in his later pea plants experiments. Nevertheless, there remain debates in statistical literature about the reasons why are Mendel's results in such a too good accordance with expected values [22, 28]. The main aim of this paper is to propose new two-stage statistical models, which are in a better accordance with Mendel's data than a classical model, where the latter considers a fixed sample size. If Mendel realized his experiments following such two-stage algorithm, which cannot be however proven, the results would purify Mendel's legacy and remove the suspicions that he modified the results. Mendel's experiments are described from a statistical point of view and his data are shown to be close to randomly generated data from the novel models. Such model is found as the most suitable, which is remarkably simpler according to the model of [28], while the new model yields only slightly weaker results. The paper also discusses Mendel's legacy from the point of view of biostatistics.

Keywords

Genetics, history of science, biostatistics, design of experiments

Correspondence to:

Jan Kalina
Institute of Computer Science CAS,
Address: Pod Vodárenskou věží 2, 182 07 Praha 8, CR
E-mail: kalina@cs.cas.cz

EJBI 2016; 12(2):en20-en26
received: October 13, 2016
accepted: December 3, 2016
published: December 31, 2016

1 Introduction

Gregor Mendel is generally acknowledged not only as the founder genetics, but also as the author of the first mathematical result in biology. Although his education had been questioned for a long time, he was profoundly educated in botany as well as physics and in those parts of mathematics (combinatorics, probability theory) applied in his later pea plants experiments. Nevertheless, there remain debates in statistical literature about the reasons why are Mendel's results in such a too good accordance with expected values [22, 28]. The main aim of this paper is to propose new two-stage statistical models, which are in a better accordance with Mendel's data than a classical model, where the latter considers a fixed sample size. If Mendel realized his experiments following such two-stage algorithm, which cannot be however proven, the results would purify Mendel's legacy and remove the suspicions that he modified the results.

Section 2 of this paper summarizes important facts about Mendel's life. His pea plants experiments are overviewed in Section 3. Also the founder of the mathematical statistics at the beginning of the 20th century were involved in their first interpretations, which is recalled in Section refkap:odpor. A statistical section 5 with an original analysis of Mendel's data is motivated by an attempt to find such design of experiments, which would be in a good accordance with Mendel's data. A newly proposed two-stage model is remarkably simpler according to the model of [28], while the new model yields only slightly weaker results. Finally, the paper also discusses Mendel's legacy from the point of view of biostatistics.

2 Mendel's biography

Gregor Mendel (20.7. 1822 – 6.1. 1884) is justly acknowledged as the founder of genetics and one of the most important biologists of all times. This sections describes Mendel's life in a much shorter way than in available monographs [11, 16, 25], but we do not neglect that Mendel acquired a profound education in mathematics and physics. It is necessary to point out in connection with Mendel's CV that prejudices against Mendel still sur-

vive in lay public or in popularization works (e.g. [18]), questioning his education or purpose of his experiments. No discussion is devoted to such prejudices here, because they have been already disproved by a series of arguments standing of proven facts or following from a historical context.

Mendel was born as Johann Mendel in a German-speaking Roman Catholic family in today's Silesian village Hynčice and was baptized in the church of St. Peter and Paul in Dolní Vražné, while both villages are nowadays part of the municipality Vražné. His parents were poor peasants and his father devoted himself to breeding fruit trees. Jan Schreiber awakened a deep interest in science education in the young Mendel. Schreiber was not only a priest in Dolní Vražné but also Mendel's teacher in Hynčice, where he taught natural science according to the spirit of Comenius. Mendel continued his studies in Lipník nad Bečvou, later in Opava at a secondary school oriented on science and finally in Olomouc at the Philosophical Institute of the university.

In 1843, Mendel joined St. Thomas' Augustinian Abbey in Old Brno, where he was ordained as a priest in 1847 and accepted a monastic name Gregor. The abbey can be denoted as a progressive education center filled with enlightenment thoughts, where monks devoted themselves to science and humanity. This was the vision of the abbot Cyril Napp (1792–1867), who was a renowned expert on breeding fruit trees himself and aimed at finding young monks with an interest science. The abbey possessed large fields and pastures, where the monks performed also sheep breeding experiments.

Figure 1: Gregor Mendel (1822–1884).

Mendel needed to pass a rigorous exam in order to promote from his position of a substitute teacher to the fully qualified one. Therefore, he went to Vienna in 1850 to undertake the exam in physics and botany. The head of the commission for the physics part of the exam was Andreas von Baumgartner and one of the members was Christian Doppler (1803–1853). Mendel was able to pass only because Baumgartner was a man of broader knowledge who preferred logical thinking to memorizing [1, 16]. However, members of the other commissions were rather pedants, which had the consequence of Mendel's failing at the whole exam. This was perhaps influenced by his being a self-learner. At any case, Napp sent a letter to Baumgartner with a question about the course of the exam. He obtained the response that Mendel made a convincing impression and as a self-learner showed his talent [1]. Then, Napp sent Mendel to study in Vienna. Mendel spent two years there (1851–1853) in a newly created scientifically oriented study program.

In Vienna, Mendel's teachers were the leading personalities and scientists in botany, physics and geology. Thus, his intellectual horizon could become much wider. He studied also at Doppler, who was the director of the newly established Institute of Experimental Physics. Mendel became acquainted in combinatorics [11], particularly with permutations and combinations in lectures in physics (and especially in meteorology). He learned also basics of probability theory and simple (but at that time not yet formalized) principles of statistical thinking [2, 5]. Mendel's second attempt for the rigorous exam in 1856 ended without success again, particularly because of botany. There are good reasons to believe however than Mendel knew the most recent scientific results in botany better than his too conservative examiners [25].

After returning to Brno, Mendel could continue teaching botany and mainly physics at various schools, while he was known as an excellent, enthusiastic and comprehensible teacher. At the same time, he could develop an intensive program on experimental plant breeding with the aim to explain laws of origin and development of hybrids. The abbot Napp, this forgotten hero in the history of genetics, let an expensive greenhouse be built for Mendel in the garden of the abbey in 1854. Mendel realized his experiments intensively here in 1856–1863; they will be described later in Section 3. Apparently, Mendel designed a detailed plan of the experiments already in Vienna [24]. He was also acknowledged as the qualified teacher even without the exam in 1856.

The modest and introvert Mendel did not cause any controversies in the abbey, he followed liberal religious stands [10] and did not even participate in the national disputes among Czechs and Germans. Both groups jointly proposed him to become the new abbot after Napp's death in 1867. As the abbot, Mendel tenaciously defended interests of the abbey against pressures from the anticlerical government.

Some sources [11, 26] claim that Mendel in his old age had no more time for pea experiments, which thus remained unfinished, and that he was isolated, lonesome and bitter and lost his personal prestige. Such sources seem to be derived from the very first Mendel's biographies (e.g. [12]), which were focused mainly on the biological problems and ignored a broader context. On the other hand, other sources [10] claimed Mendel to acquire a remarkable reputation in the society as well as inner peace. Because he himself considered the experiments to be finished, he was able to devote much time to meteorology, beekeeping and experiments with breeding decorative plants and fruit trees [21, 25]. Finally, he passed away in 1884 in connection to his chronic kidney disease.

3 Mendel's experiments

Mendel spent 8 years with intensive experiments with pea plants. They were apparently performed with the aim to empirically verify principles of heredity, while he was able to understand them theoretically in a correct way already at that time from the points of view of both biology and mathematics [3]. The experiments were performed according to a clear vision and following a theory, which was elaborated prior to the experiments. Mendel so much loved the experiments that he would not allow them to any his assistant [21] and the big greenhouse served only Mendel and only pea plants. He started by selecting varieties suitable for the experiments and invented also a new methodology for heredity research. In very simple conditions, he was able to cross and investigate over 12 000 pea plants. Mendel was able to gather a huge amount of experimental data thanks to his (we can say mathematical) thoroughness and precision [10].

Mendel presented his selected results to a community of experts on botany and breeding in Brno twice during 1865. He made the impression of a great experimenter, who was perceived as unfortunately spoiled by mathematics, because he devoted the talk to randomness and probability evaluations. Even in spite of it, he was offered a possibility to publish his summarizing results. Thus, his only paper on heredity [19] was published in 1866. It did not however arouse any attention and was practically ignored so that the core his discovery was lost for one whole generation of Mendel's contemporaries.

The text of the whole paper [19], which was also repeatedly translated to various languages, has the experiments and measured data as its central topic. Mendel does not formulate genetic laws in a general form. This is why the paper gives a rather complicated impression at first glance. At second glance, or during repeated reading, a today's reader may already understand Mendel's thoughts, which is however to a large extent a consequence of our understanding of elementary genetic laws. In Mendel's times, the paper must made a blurry impression. Mendel denoted variables by letters like it is common in contemporary algebra. However, when he considered counts and probabilities in the paper, he did not use contemporary mathematical terminology, which was not defined at that time. He either defined his own concepts or tried to circumscribe them in an idiosyncratic fashion.

From the mathematical point of view, Mendel identified that the ratio 3:1 of e.g. green and yellow unripe legumes in reality corresponds to the ratio 1:2:1, where the green color is dominant and the ratio corresponds to green legumes with genotype AA, green Aa and yellow aa, respectively. Perhaps it is the very grasping the mathematical structure of the results and these abstract thoughts that reveal Mendel most clearly as a man of genius.

Mendel's results in meteorology, in which he published 9 papers, represent an independent topic. Indeed, he was far ahead his time also in meteorology [20]. His description and explanation of causes of strong wind effects, which also destroyed his greenhouse, is remarkable. Towards the end of his life, Mendel devoted much time to a thorough weather measurement and his hand-written reports from 1879–1883 constitute the basis of regular meteorological observations in Brno.

4 Statistics as a deteriorating circumstance

Mendel's only paper on heredity was rediscovered only in 1900 in a way which was denoted as a fairy-tale [5]. Soon, Mendel became a target of questioning results, pseudoscientific interpretations as well as fanatic attacks against his personal reputation. Thus, he was to a large extent considered as a controversial or naïve amateur, dilettante or uneducated person, who was able to come to important discoveries only by a mere chance and good luck behind the walls of the abbey. Also important statisticians of the beginnings of the 20th century contributed to such assessment of Mendel's legacy.

One of main Mendel's critics was Karl Pearson (1857–1936), who was the first professor of statistics in the world [15]. In the Biometrika journal, he criticized Mendel. He considered his results absurd and the statistical arguments half-bogus [4]. He founded the statistical school of biometricians, who attacked Mendel's results for being extremely close to expected values by means of the Pearson's χ^2 goodness-of-fit test. While biometricians believed that heredity (i.e. genetic variability) is continuous, another school denoted as Mendelists opposed them with the idea of a discrete heredity. William Bateson (1861–1926), their leading personality and biologist, believed Mendel's statistical results, but he did not understand the statistical arguments and even considered statistics in biology to be senseless and useless [17].

The controversy of both schools was ended by Ronald A. Fisher, a leading biostatistician and biologist, who confirmed Mendel's idea of a discrete heredity. Although the result was a practical victory of Mendelists, Fisher was at the same time able to reconcile both groups. Fisher admired Mendel with humility [5, 14] and also used his data to illustrate some of the novel statistical methods (e.g. proposed in [7]), although he was deeply convinced about Mendel's falsifying the data.

Let us remark that Mendel's legacy was not allowed to be commemorated in Czechoslovakia after 1948 until 1960s [32], not only because of his clerical background, but also with the justification that Mendel exploited statistical reasoning [30].

Currently, there continue debates of experts attempting to explain why are Mendel's results biased towards expected values. Various scientific papers (e.g. [21, 27, 28]) have recently tried to rehabilitate Mendel. One of the arguments is based on the idea that he could have performed the experiments according to a more complex design. So far, there have appeared no remarkable arguments against

it. The following section proposes new possible models. Numerical simulations are used to find out if Mendel's original data are in accordance with these newly proposed designs.

5 Two-stage models for Mendel's experiments

In literature, there have been intensive discussions attempting to find explanation for a too good accordance between Mendel's data and expected values. In such context, expected values are those which would have been obtained in ideal experiments without any nuisance external effects under the assumption that the randomness may not influence the results in any extreme manner [27, 29]. Such explanation remains however unknown also as a consequence of lacking knowledge of the detailed organization of all Mendel's experiments as well as of lacking preliminary results. Statistical attempts for such explanation include proposed two-stage models of [13, 28], which assume Mendel to decide for more experiments if the results do not yield a sufficiently remarkable confirmation of the theoretical model. There is no indication that Mendel used a two-stage design but this is not impossible. One more reason for it is a lack of outlying measurements (outliers) in the data [22]. In this section, a new two-stage model is proposed and compared with the model of [28] by means of numerical simulations. Also, a method for finding an optimal constant, on which the model depends, is proposed.

We work with results of Mendel's 84 experiments. The number of pea plants, for which Mendel collected the data, ranges between 19 and 8023 in various experiments. Particularly, we explain the two-stage models on one experiments devoted to flower color. He hypothesized that purple and white flower appear in the population exactly in the ratio 3:1. This corresponds to the probability $\pi_0 = 3/4$ for purple. Mendel plants $n = 8023$ pea plants. Consequently, he observed purple flowers (dominant trait) exactly for $X = 6022$ plants. We consider X as a realization of a random variable with a binomial distribution $\mathsf{Bi}(n, \pi_0)$, where the appearance of the dominant trait is denoted as a success. The expected value for the number of successes equals $n\pi_0 = 6017.25$. Other Mendel's experiments are considered in an analogous way leading to binomial distributions with different values of π_0, namely 1/2, 2/3, or 3/4.

Pires and Branco [28] claimed that Mendel could have used a two-stage design based on the χ^2 test for the binomial distribution. The whole approach will be described in a rather more complicated way now in order to allow for comparisons of different models. Let us first introduce the notation (X, n, π_0) for the triple of values, where X is a random variable following the binomial distribution $\mathsf{Bi}(n, \pi_0)$. The model (design) will be described by Algorithm 1, where $\chi^2(X, n, \pi_0)$ denotes the value of the χ^2 statistic for testing the null hypothesis $H_0 : \pi = \pi_0$ against the alternative hypothesis $H_1 : \pi \neq \pi_0$, i.e.

$$\chi^2 = \frac{(X - n\pi_0)^2}{n\pi_0(1 - \pi_0)}. \tag{1}$$

Algorithm 1 The two-stage model of [28].

Input: $n_1 \in \mathbb{N}$, $n_2 \in \mathbb{N}$, $c_1 \in [0, 1]$, $\pi_0 \in (0, 1)$
Output: Number of observations n, number of successes X, corresponding p-value p

```
Perform n_1 measurements (i.e. on n_1 plants)
X_1 := number of successes among n_1 measurements
p_1 := p-value corresponding to χ²(n_1, X_1, π_0)
if p_1 ≥ c_1 then
    n := n_1, X := X_1, p := p_1
else
    Perform n_2 measurements and denote the number of successes as X_2
    p_2 := p-value corresponding to χ²(n_2, X_2, π_0)
    n := n_i, X := X_i and p := p_i, where i = arg max p_i over i ∈ {1, 2}
end if
```

Such approach, when the measurement is performed either once or twice, is denoted as two-stage. The classical approach, performing all measurements at once with a fixed sample size, is denoted as one-stage. We already proposed an alternative two-stage algorithm in the paper [13], where however only values from the second block of measurements is used (i.e. if the second block is performed), ignoring the whole first block. The new model is proposed in Algorithm 2, exploiting the notation

$$Z = \frac{|X - n\pi_0|}{n}. \tag{2}$$

Algorithm 2 New two-stage model.

Input: $n_1 \in \mathbb{N}$, $n_2 \in \mathbb{N}$, $c_2 \geq 0$
Output: Number of observations n, number of successes X, corresponding p-value p

```
Perform n_1 measurements
X_1 := number of successes among n_1 measurements
p_1 := p-value of the χ² test for (n_1, X_1, π_0)
if Z < c_2 then
    n := n_1, X := X_1, p := p_1
else
    Perform n_2 measurements and denote the number of successes as X_2
    p_2 := p-value of the χ² test for (n_2, X_2, π_0)
    n := n_i, X := X_i and p := p_i, where i = arg min Z_i over i ∈ {1, 2}
end if
```

Performing the experiments using Algorithm 2 neither require the computation of the χ^2 statistic nor of the p-value. Because both were unknown in Mendel's times, the approach based on Algorithm 1 seems much more intuitive. We included computations of the p-values only artificially in order to allow comparisons which of the models

is the best for explaining Mendel's data. We will generate random data from Algorithm 1 with various c_1 and from Algorithm 2 with various c_2. Simulated data will be generated from the binomial distribution with the same values of n and π_0 as in Mendel's experiments.

It is now necessary to measure the distance between Mendel's data and simulated data from Algorithms 1 and 2. Individual p-values are random variables and the aim is not to compare them themselves, but the whole set of 84 p-values. It is worth noting that values of χ^2 or Z themselves depend on the sample sizes and thus are not suitable as distance measures. Indeed, the task is to find a suitable measure of distance between two distributions or a loss function expressing the loss of a particular two-stage model compared to Mendel's observed data. We consider the measures

$$M_1 = \sum_{i=1}^{84} |p^1_{(i)} - p^M_{(i)}| \quad \text{and} \quad M_2 = \sum_{i=1}^{84} |p^2_{(i)} - p^M_{(i)}|, \quad (3)$$

to be suitable, where

- $\left(p^M_1, \ldots, p^M_{84}\right)^T$ = vector of p-values of the χ^2 test for 84 Mendel's experiments,
- $\left(p^j_1, \ldots, p^j_{84}\right)^T$ = vector of p-values of the χ^2 test for randomly generated data from Algorithm j, where $j \in \{1, 2\}$ and c_j is given.
- $p^M_{(1)} \leq \ldots \leq p^M_{(84)}$ = arranged values $p^M_1, \ldots, p^M_{84}$,
- $p^j_{(1)} \leq \ldots \leq p^j_{(84)}$ = arranged values $p^j_1, \ldots, p^j_{84}$ for $j \in \{1, 2\}$.

Naturally, M_1 depends on selected c_1 and M_2 depends on c_2. We will perform the random generation 1000-times and averaged values of M_1 and M_2 will be considered.

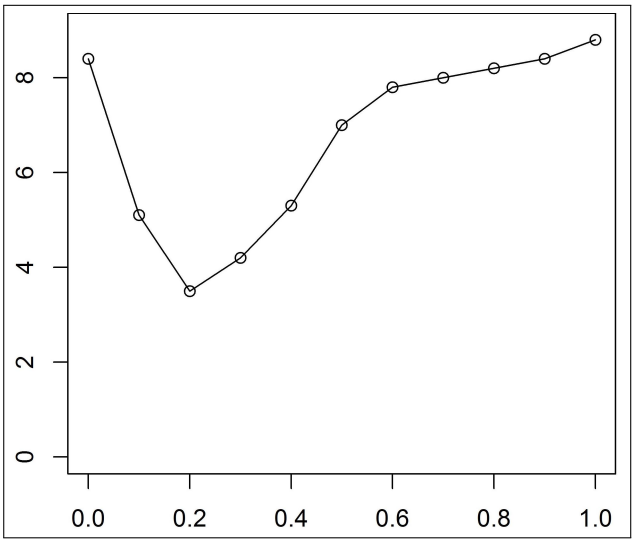

Figure 2: Comparison of data generated from Algorithm 1 with Mendel's data, depending on the choice of c_1. The comparison is performed by means of the measure M_1.

Figure 2 shows computed values of the measure M_1 depending on the choice of c_1. The values $c_1 = 0$ corresponds to the one-stage model. If $c_1 = 1$, the maximum of both p-values is considered. We can see, and it is revealed also by other more detailed analysis, that Algorithm 1 is the closest to Mendel's data for $c = 0.2$, which corresponds to the optimal value $c = 0.201$ of [28]. The improvement compared to the one-stage model is remarkable.

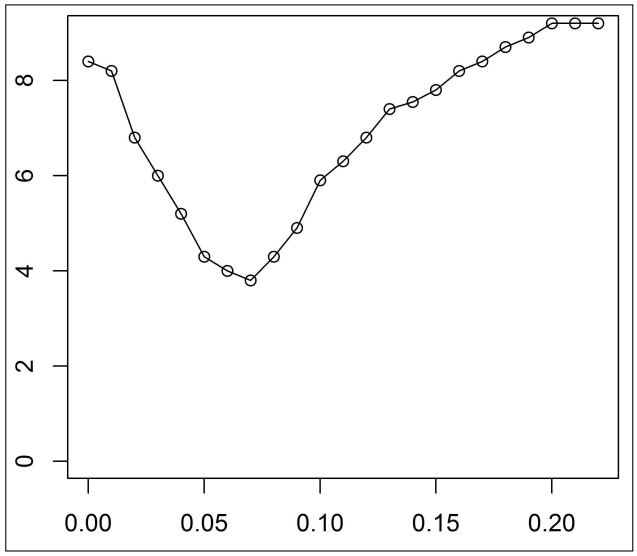

Figure 3: Comparison of data generated from Algorithm 2 with Mendel's data, depending on the choice of c_2. The comparison is performed by means of the measure M_2.

Figure 3 shows computed values of the measure M_2 depending on the choice of c_2. The interpretation of the figure is different from Figure 2. The value $c_2 = 0$ now corresponds to the minimum of two values of Z, while a sufficiently large c_2 (not limited from above) corresponds to the one-stage model. The optimal value of c_2 in Algorithm 2 is $c_2 = 0.08$.

Table 1: Added value of two-stage models compared to the one-stage model. The numbers are values of a distance measure between the two-stage model and Mendel's data divided by the distance measure between two one-stage model and Mendel's data.

Algorithm	Ratio
1	0.42
2	0.45
3	0.51
4	0.70

We performed additional computations for other two-stage models, which lead only a slight improvement compared to the one-stage model. These include Algorithm 3, which differs from Algorithm 2 only in $X := X_1 + X_2$ in case that $Z < c_2$. Algorithm 4 differs from Algorithm 2 only in $X := X_2$ in case that $Z < c_2$. Table 1 shows the relative ratio of improvement of individual models compared to the one-stage model. For example, Algorithm 1 allows to reach (with the optimal value of c_1) only 42 % of the value of M_1 attainable with the one-stage model (with

$c = 0$). Very similar results are obtained if a quadratic distance measure between two vectors instead of (3) is used.

On the whole, the computations and figures show that the idea of a two-stage approach is definitely not meaningless. Algorithm 1 corresponds to Mendel's data better than Algorithm 2, but this superiority is practically negligible. On the other hand, Algorithm 2 gives a more likely impression for its simplicity from the points of view of interpretation and computation. Algorithms 3 and 4 are less reliable for describing Mendel's data.

6 Mendel's legacy from the point of view of biostatistics

Mendel was the first to explain the substance of heredity. He developed the methodology for the study of heredity, which has been used until now, and strongly influenced current genetic engineering [32]. He belongs also to the most important theoretical biologists as the founder of genetics. At his time, nobody would expect that general genetic laws could be derived by means of (possibly large) experiments on an only plant. Also pea plants were too ordinary for such far-reaching experiments. Nowadays, genetic laws are denoted by Mendel's name and their correctness was theoretically proven after the discovery of DNA. Mendel has been also acknowledged as the founder of bioinformatics thanks to his discovery of the substance of hereditary bioinformation. His understanding the gene as an algebraic unit truly represents a jump to the 21th century [30].

Mendel was also the first to exploit combinatorics, probability theory and mathematics in general in biology. This was a revolutionary step already by understanding very significance and influence of randomness in heredity. He explained that randomness is manifested as a discrete variable, which allowed him to derive probabilities for the genotype and phenotype of the offsprings. He had contributed to constructing modern biology on statistical thinking even before mathematical statistics started its existence. Botany experts of that time perceived it as a contamination of their science, but we can view statistics as an inseparable part of current biological or biomedical research. Mendel used only a naïve definition of probability and intuitive inductive thinking [5], which was sufficient for the particular statistical comparison of his results with expected values. In spite of his contribution to the development of biostatistics, Mendel is not generally acknowledged as its founder.

Mendel influenced also sophisticated design of experiments, which belongs to statistics as its integral part. He organized the experiments in an unprecedented way although the design seemed obscure to his contemporaries. The design required to consider and evaluate an enormous number of plants within the experiments. Fisher [6] derived statistical formulas for their analysis and we nowadays understand the whole procedure, which is now known as factorial design, as natural, intuitive and standard. It is worth mentioning that the statistical concept of factorial design comes from Mendel's notion of factors (in German *die Faktoren*) for genes. This concept was later adopted by Fisher for variables measured in any (not only genetic) experiment.

Mendel is also denoted as the progenitor of Mendelian randomization, which obtain increasing popularity in analytical epidemiological studies [23, 31]. If the influence of various factors (affected by environment and/or genetics) on the treatment of patients is investigated, randomized control studies have to face a number of practical limitations. It is for example impossible to perform a randomization of the genotype, which is given by a random combination of genes of a particular patient's parents, but is already fixed at the time of the study, because the randomization of alleles was performed by the nature itself at the time of conception. Another hardly imaginable requirement on a strict randomization would be to push a patient to become a smoker or alcoholic for the purpose of the study.

The Mendelian randomization is a procedure stemming directly from the second Mendel's law, which is able to replace the machinery of randomized clinical trials. From the statistical point of view, the method represents a correction for a systematic error (confounding) in the design of the study. It is namely common that patients selected for a clinical study have a better (or in the contrary worse) prognosis compared to patients treated with conventional methods. The principle of the method is statistical and consists in approaching a particular genetic variant as a natural instrumental variable, while hypothesis testing or estimating the effect is performed by means of the instrumental variable estimator, which is a popular procedure mainly in econometrics.

It will be never more possible to prove without doubt if Mendel modified his experimental data and performed an intentional scientific misconduct or not. This cannot be decided even by means of a statistical analysis. This paper however contributes to the discussions if the data could really arise without his falsifications. Numerous works in scientific journals have in recent years attempted to rehabilitate Mendel from the accusations of modifying or even creating the results. Their arguments come from different fields of genetics, breeding, history, ethics and even psychology, philosophy or theology [21, 24]. According to such recent arguments, our whole knowledge about Mendel shows his intentional modifications of results to be highly improbable [21]. In fact, it were rather Mendel's critics who performed experiments violating ethics or credibility principles, such as eugenicists or the Soviet biologist T.D. Lysenko (1898–1976) [8]. Thus it seems that Gregor Mendel, the respectable and exemplar priest, can finally come out from all the controversies and accusations as a moral authority and man of the noblest character [10].

Acknowledgement

The author is thankful to two anonymous referees for interesting suggestions.

References

[1] Corcos, A.F., Monaghan, F.V.: *Gregor Mendel's experiments on plant hybrids: A guided study.* Rutgers University Press, New Brunswick, 1993.

[2] Edelson, E.: *Gregor Mendel and the roots of genetics.* Oxford University Press, New York, 1999.

[3] Fairbanks, D.J., Rytting, B.: *Mendelian controversies: A botanical and historical review.* American Journal of Botany **88** (2001), 737 – 752.

[4] Fisher, R.A.: *Statistical methods in genetics.* Heredity **6** (1952), 1 – 12.

[5] Fisher, R.A.: *Has Mendel's work been rediscovered?* Annals of Science **1** (1936), 115 – 137.

[6] Fisher, R.A.: *The design of experiments.* Oliver & Boyd, Edinburgh, 1935.

[7] Fisher, R.A.: *The correlation between relatives on the supposition of Mendelian inheritance.* Transactions of the Royal Society of Edinburgh **52** (1918), 399 – 433.

[8] Flégr, J.: *A different view on Lysenkoism.* Vesmír **78** (1999), 16 – 21. In Czech.

[9] Franklin, A., Edwards, A.W.F., Fairbanks, D.J., Hartl, D.L., Seidenfeld, T.: *Ending the Mendel-Fisher Controversy.* University of Pittsburgh Press, Pittsburgh, 2008.

[10] Gustaffson, A.: *The life of Gregor Johann Mendel —Tragic or not?* Hereditas **62** (1969), 239 – 258.

[11] Henig, R.M.: *The monk in the garden: The lost and found genius of Gregor Mendel, the father of genetics.* Houghton Mifflin Harcourt, Boston, 2000.

[12] Iltis, H.: *Johann Gregor Mendel, Leben, Werk und Wirkung.* Springer, Berlin, 1924.

[13] Kalina, J.: *Gregor Mendel's experiments and their statistical evaluation.* Acta Musei Moraviae, Scientiae Biologicae **99** (2014), 87 – 99.

[14] Kalina, J.: *Ronald Fisher, the father of biostatistics.* Pokroky matematiky, fyziky a astronomie **57** (2012), 186 – 190. In Czech.

[15] Kalina, J.: *On the anniversary of the death of the founder of biometrics.* Biologie-Chemie-Zeměpis **20** (2011), 123 – 127. In Czech.

[16] Klein, J., Klein, N.: *Solitude of a humble genius Gregor Johann Mendel.* Volume 1. Springer, Berlin, 2013.

[17] Malý, M.: *Biometrika one hundred years old.* Informační bulletin České statistické společnosti **14** (2003), 4 – 9. In Czech.

[18] Mawer, S.: *Mendel's dwarf.* Kniha Zlín, Zlín, 2010. In Czech.

[19] Mendel, G.: *Versuche über Pflanzen-Hybriden. Verhandlungen des Naturforschenden Vereines in Brünn* **4** (1866), 3 – 47. In German.

[20] Munzar, J.: *Gregor Mendel—meteorologist.* Československý časopis pro fyziku A **31** (1981), 63 – 67. In Czech.

[21] Nissani, M.: *Psychological, historical, and ethical reflections on the Mendelian paradox.* Perspectives in Biology and Medicine **37** (1994), 182 – 196.

[22] Novitski, E.: *On Fisher's criticism of Mendel's results with the garden pea.* Genetics **166** (2004), 1133 – 1136.

[23] Novotný, L., Bencko, V.: *Genotype–disease association and possibility to reveal environmentally modifiable disease causes: The use of Mendelian randomization principle.* Časopis lékař ˚u českých **146** (2007), 343 – 350. In czech.

[24] Orel, V.: *Science studies in evaluating Mendel's research.* Informační listy Genetické společnosti Gregora Mendela **32**, 2007, 20 – 27. In Czech.

[25] Orel, V.: *Gregor Mendel and beginnings of genetics.* Academia, Praha, 2003. In Czech.

[26] Petřík, M.: *Saints and witnesses: Gregor Johann Mendel.* TV document, 2008. Available from: http://www.ceskatelevize.cz/porady/10123417216-svetci-a-svedci/208562211000012-gregor-johann-mendel. In Czech.

[27] Piegorsch, W.W.: *Fisher's contributions to genetics and heredity, with special emphasis on the Gregor Mendel controversy.* Biometrics **46** (1990), 915 – 924.

[28] Pires, A.M., Branco, J.A.: *A statistical model to explain the Mendel-Fisher controversy.* Statistical Science **25** (2010), 545 – 565.

[29] Sekerák, J.: *Mendel in a black box.* Moravské zemské muzeum, Brno, 2008. In Czech.

[30] Sekerák, J.: *Problem of a philosophical interpretation of a scientific text: G.J. Mendel.* Dissertation thesis, Masarykova Univerzita v Brně, Brno, 2007. In Czech.

[31] Smith, G.D., Ebrahim, S.: *Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease?* International Journal of Epidemiology **32** (2003), 1 – 22.

[32] Vondrejs, V.: *Questions around genetic engineering.* Academia, Praha, 2010. In Czech.

Students' Behavior Related to Oral Health

Tanja Marković[1]

[1] Faculty of Medicine, University of Montenegro, Podgorica, Montenegro

Abstract

Introduction: Oral health significantly affects the quality of life. Behavior is a very important determinant of oral health, which is related to oral hygiene, diet, regular visits to the dentist and smoking. Students are interesting research group to study an oral-helath behavior, particularly dental students, because of their knowledge and attitudes about oral health greatly affect the proceedings in their further work which is important for the health of the population.
Objective: The objective was to evaluate the oral health behavior of dental students of the Medical Faculty in Podgorica and the Faculty of Political Science in Podgorica.
Materials and Methods: The study included 125 students. The survey instrument was a questionnaire of closed type, containing questions about oral hygiene, visits to the dentist, as well as questions about nutrition and consumption of tobacco.
Results: The study showed that students take care of their oral health; 75.2% brush their teeth 2 to 3 times a day, and 54.4% goes to examinations every 3 - 6 months. Results also indicated that there are certain problems that require health-educational measures, such as a lack of knowledge about fluoride as an element of dental caries prevention, as well as the insufficient use of additional funds for oral hygiene, such as dental floss and interdental brushes.
Conclusion: The study showed that students take care of their oral health, but there is a need for continuous education programs on the importance and protection of oral health. Dental students have shown a bit better results.

Keywords

Oral health, behavior, students, dental hygiene, dentistry

Correspondence to:

Tanja Marković
Faculty od Medicine, University of Montenegro, Podgorica
Address: Jola Piletića 8, Nikišić, Montenegro
E-mail: tanja.markovic88@hotmail.com

EJBI 2016; 12(2):en27-en31
received: April 2, 2016
accepted: October 8, 2016
published: December 31, 2016

1 Introduction

Oral health significantly affects the quality of life. The presence of oral disease reduces the feeling of self-esteem, affect the diet, as well as the ability to communicate and the overall health both in childhood and in the elderly [1]. Today, aesthetic role of teeth is also very important. Considering that the main cause of dental and periodontal disease is dental plaque, oral hygiene, which involves the complete removal of it, may be considered the dominant determinant of oral health [2].

Behavior is a very important determinant of oral health [3, 4]. When talking about behavior which is important for oral health, we primarily refer to the maintenance of oral hygiene, diet and regular dental visits.

Adequate oral hygiene is the most efficient and simplest method of preventing tooth decay. Removing plaque from all tooth surfaces requires the use of additional funds for oral hygiene, in addition to a toothbrush and toothpaste. These are different interdental brushes, gingival toothbrushes, dental floss, mouthwash solutions, etc. [5, 6].

Smoking, the most common form of helath-risk behavior, contributes to the development of periodontal diseases and causes greater resorption of alveolar bone [7].

A very important factor in maintaining oral health is a proper diet [8]. Frequent consumption of refined carbohydrates influences the appearance of cavities [9]. Dental plaque bacteria's degradation of carbohydrates release acids that favor the demineralization of tooth enamel, a process that marks the beginning of the creation of the carious lesion.

Regular check-ups at the dentist are also very important, becuose the dentist can detect initial carious lesions and react promptly.

The use of fluoride is an important preventive measure. Fluoride prevents initial demineralization of tooth enamel and have the inhibitory effect on the metabolism of cariogenic bacteria [10].

In addition to the measures already described, the health educational work with the population has an im-

Table 1: Frequency of toothbrushing in relation to faculty and sex.

Frequency of toothbrushing	Dentistry		Faculty of Political Science	
	Male	Female	Male	Female
More than 3 times	5 (15.6%)	5 (12.2%)	1 (10%)	13 (31%)
2 – 3 times	26 (81.3%)	36 (87.8%)	7 (70%)	25 (59.5%)
Once a day	1 (3.2%)	/	2 (20%)	4 (9.5%)
Never	/	/	/	/
Total	32 (100%)	41 (100%)	10 (100%)	42 (100%)

Table 2: Usage of various types of toothbrushes and frequency of changing toothbrushes in correlation with the faculty.

Types of toothbrushes	Frequency of changing toothbrushes							
	Dentistry				Faculty of Political Science			
	Monthly	When the fibres change	2-6 months	Doesn't follow	Monthly	When the fibres change	2-6 months	Doesn't follow
Soft or ultra-soft	12 (24%)	24 (48%)	11 (22%)	3 (6%)	4 (40%)	4 (40%)	2 (20%)	
Medium	3 (13.6%)	10 (45.5%)	9 (40.9%)	/	12 (32.4%)	14 (47.8%)	10 (27%)	1 (2.7%)
Hard	/	1 (100%)	/	/	3 (60%)	1 (20%)	1 (20%)	
Doesn't use toothbrush	/	/	/	/	/	/	/	/
Total	15	35	20	3	19	19	13	1

portant role in maintaining and improving oral health [5]. Therefore, dental health education should be part of general education [2].

Application of fluoride in addition to public health care programs in the Nordic countries has caused a significant drop in caries prevalence [11].

Students, as part of the population that has a significant role in the development of any society, are a very interesting research group to study the oral health behavior.

The students of dentistry are very important because their knowledge and attitudes about oral health and preventive measures greatly affect the proceedings in their further work which is of great importance for the health of the population.

Numerous studies have shown that students of dentistry have a higher level of knowledge and take better care of oral health compared to students of other faculties [12, 13]. It was shown also that their behavior in relation to oral health changes during the years of study; senior students demonstrate better results [14].

In Montenegro, studies which should show oral health behavior of students had never been implemented. Similar studies done in school children in Montenegro have indicated the need for continuous educational programs on oral health and hygiene of the mouth and teeth [3, 15]. The aim of the research is to investigate the oral health behavior of students of Dental Medicine and the Faculty of Political Science in Podgorica.

2 Material and Methods

The survey was conducted from February to April 2016, at Department of Dentistry, Medical Faculty in Podgorica and the departments of Social Work and International Relations, Faculty of Political Sciences in Podgorica. The study included 73 dental students and 52 students of the Faculty of Political Science. All students have voluntarily agreed to participate in the study after they have been explained about the purpose of the test, and the participation rate was 100%. Respondents answered all the questions, which means that the response rate was 100%. The survey instrument was questioner which was created following the examples on similar research conducted in Turkey and Greece [16, 19].

The questionnaire consisted of 17 questions. The first group of questions has been related to general information about the respondents - gender, study program and year of study they attend. The next group of questions has been related to oral hygiene: how often do they brush teeth, which means for oral hygiene they use, what kind of toothbrush they use and how often they change it.

There was also a set of questions, related to the dental visit (how often they go to the dentist, which are the most common reasons for visits) as well as the role of dentists in health education. Respondents were also asked about the prevention of oral diseases and the availability of information about oral health, as well as about nutrition and tobacco use.

For statistical analysis were used basic methods of descriptive and inductive statistics. As statistically significant p-values were taken those less than 5% ($p < 0.05$).

3 Results

The study included 125 participants - 73 students of the Faculty of Medicine, department Dentistry and 52 students of the Faculty of Political Sciences. Among the respondents there was 66.4% females (Figure 1).

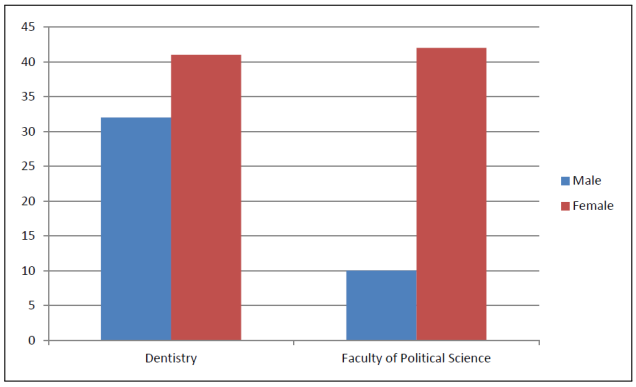

Figure 1: Distribution of the number of respondents by gender and faculty.

The largest number of respondents (75.2%) brush their teeth 2 to 3 times a day (Table 1). Of the total number of future dentists, 13.7% of them brush their teeth more than three times a day, while among the students of Political Science, there are slightly more of them (26.9%).

When it comes to the type of toothbrush the respondents use, it was observed that there is a correlation between the use of certain toothbrushes and faculty they attend (Pearson correlation coefficient r = 0.485, p <0.001). Unlike the kind of toothbrush, statistical analysis shows that there is no correlation between the frequency of changing toothbrushes and faculty students attend (r = 0.243, p> 0.05), and no correlation with sex (r = 0.138, p> 0 05). Most students (43.2%) change the brush when recognizing that the fibers are changed (Table 2).

The data show that there is a statistically significant difference in the use of toothpaste containing fluoride between the students of the Faculty of Political Science and Dentistry students (r = 0.704, p <0.001). Of the total number of students of Political Science 71.2% do not know whether their toothpaste contains fluoride, while the majority of future dentists (87.7%) use toothpaste with fluoride, as shown in Table 3.

Statistical analysis of data showing the frequency of use of additional funds for oral hygiene, such as dental floss and interdental brushes, indicates that there are differences in their use among future dentists and students of Political Science (r = 0.264, p <0.05). Of the total number of dental students, 31.5% of them use additional funds, while 13.5% of students of political science is not informed of the existence of an additional funds for the oral hygiene, as shown in Table 4.

4 Discussion

Research has shown that students take care about oral health, but there are problems that require a public health intervention. As expected, the dental students show better results. A similar study conducted in Turkey showed that only 12% of the tested students regularly go to dental check-ups, while Montenegrin students show better results - 54.4% of respondents every 3 - 6 months visit the dentist, which can be a result of the fact that our students don't pay for dental services [16].

Research has shown that students of the Faculty of Political Science do not have enough knowledge about fluoride, as an important element of dental caries prevention; 71.2% of them don't know if their toothpaste contains fluoride, while expected, the majority of future dentists (87.7%) use toothpaste with fluoride. Half of Turkish students (52.7%) use toothpaste with fluoride [16].

The study shows that 57.6% of respondents had training on proper tooth brushing, which are similar results as in Turkey (57.3%) [16].

Expectedly, dental students showed knowledge of the proper choice of toothbrush - 68.5% use a soft or ultra-soft brush, which is today considered to be optimal for maintaining oral hygiene, while most students of political science (71.2%) use medium brush. The survey conducted in Hiroshima, which compared oral-health behavior between dental and civil engineering students, also showed that future dentists have more knowledge about the proper choice of toothbrush [17]. Data showing the use of additional funds for oral hygiene are not satisfactory; only 24.8% of students use of dental floss and interdental brushes every day. This is a slightly better result than that achieved in Belgrade adolescents, of whom 13.4% use dental floss [4].

One third of future dentists (31.5%), use dental floss and interdental brushes every day, which is unexpectedly bad result.

Almost all respondents (84.8%) agree that the public in Montenegro is not sufficiently informed about the importance of oral health and prevention of oral diseases, which indicate a need for oral- health programs.

A similar research in Montenegro, conducted among school children, also indicated a need for education on the conservation of oral health [3, 15].

A similar survey of oral health behavior in Foca, which included schoolchildren, shows that respondents most often go to the dentist when they have toothache (51%), while our students expected, due to the difference in age, show better results - the most common reason for visiting the dentist (69.6%) are regular checks [18].

The survey of oral health behavior of medical students in Greece shows that 51.25% of respondents go to the dentist when they have toothache, which are inferior results compared to the our study [19]. Future dentists in Peru also show worse results, compared to our students; 28% of first year students and 6% of fifth year students go to the dentist when they have toothache [20].

Table 3: Correaltion between usage of fluoride toothpaste and faculty.

Fluoride toothpaste	Dentistry	Faculty of Political Sciences
Yes	64 (87.7%)	9 (17.3%)
No	3 (4.1%)	6 (11.5%)
Doesn't know	6 (8.2%)	37 (71.2%)
Total	73 (100%)	52 (100%)

Table 4: Distribution of usage of additional funds for oral hygiene by gender and faculty.

Usage of additional funds for oral hygiene	Dentistry		Faculty of Political Science	
	Male	Female	Male	Female
Yes, every day	8 (34.8%)	15 (65.2%)	/	8 (100%)
Yes, sometimes	24 (49%)	25 (51%)	7 (18.9%)	30 (81.1%)
Doesn't know for additional funds	/	1 (100%)	3 (42.9%)	4 (57.1%)
Total	32	41	10	42

The data showed that 79.5% of dental students does not have the problem with bleeding gums when brushing teeth, which indicates adequate maintenance of oral hygiene. Of the total number of tested medicine students in Greece, 7.25% of them have problems with bleeding gums when brushing teeth, which was better results than ours, where on the question of gingival bleeding positively responded 24.8% of students [19]. Results similar to ours were obtained in the study in Peru; 31% of first year students and 9% of fifth year students have problems with bleeding gums [20].

The results show that a relatively small number of students (14.4%) have bleached teeth, mainly at home (8%). The results of research in Turkey, show that an even smaller number of respondents (3.7%) have blaeched teeth [16].

Data on the use of tobacco showed that among students there are 21.6% of smokers. These are slightly better results compared to the testing of the student population in Kosovo, where there is 27% of smokers and in Novi Sad (26.7%), while among the tested adolescents in Belgrade there was 23.3% of smokers [4, 21, 22].

When it comes to diet, the results are relatively satisfactory; 46.4% of students consume fruit several times a day, while 32.8% of them eat vegetables once a day. However, it is worrying that 31.2% of respondents eat sweets several times during the day, which is in direct correlation with caries [9, 23].

5 Conclusion

The survey showed that respondents take care of their oral health, but there is a requirement for continuous educational programs of the importance and preservation of oral health. Dental students show better results, which is of particular importance because it is their future work of importance for the health of the population.

References

[1] Davidović B, Janković S, Ivanović D, Grujičić I. Oral health assessment among dental students. Serbian Dental Journal. 2012; 59 (3): 141-7. [Cited 2016 Jan 31]. Available from: http://scindeks-clanci.ceon.rs/data/pdf/0039-1743/2012/0039-17431203141D.pdf

[2] Matijević S. Oral hygiene as the dominant determinant of health. Acta Stomatologica Naissi. 2013; 29 (68): 1,298 to 305. [Cited 2016 Feb 2]. Available from: http://scindeks.ceon.rs/article.aspx?query=ARTAK%26AND%26oralno%2bzdravlje%2b&page=15&sort=1&stype=1&backurl=%2fSearchResults.aspx%3fquery%3dARTAK%2526AND%2526oralno%252bzdravlje%252b%26page%3d0%26sort%3d1%26stype%3d1

[3] Andelić I, Matijević S, Andelić J. The importance of oral health behaviuor of children for their oral health. Sanamed. 2015; 10 (2): 101-7. [Cited 2016 Feb 2]. Available from: http://scindeks.ceon.rs/article.aspx?query=ARTAK%26AND%26oralno%2bzdravlje%2b&page=7&sort=1&stype=1&backurl=%2fSearchResults.aspx%3fquery%3dARTAK%2526AND%2526oralno%252bzdravlje%252b%26page%3d0%26sort%3d1%26stype%3d1

[4] Lalić M, Krivokapic M, Jankovic-Letter M, Aleksić E, Gajić M, Banković D. Influence of oral health related behaviuor on oral health of adolescents in Belgrade. Serbian Dental Journal. 2013; 60 (2): 76-84. [Cited 2016 Oct 5]. Available from: http://scindeks.ceon.rs/article.aspx?query=ARTAK%26AND%26oralno%2bzdravlje%2b&page=25&sort=1&stype=1&backurl=%2fSearchResults.aspx%3fquery%3dARTAK%2526AND%2526oralno%252bzdravlje%252b%26page%3d0%26sort%3d1%26stype%3d1

[5] Ljaljević A, Matijević S, Terzic N, Andelić J, Mugoša B. Signigicance of proper oral hygiene for helath condition of mouth and teeth. Vojnosanitetski Pregled. 2012; 69 (1): 16-21. [Cited 2016 Feb 2]. Available from: http://scindeks.ceon.rs/article.aspx?query=ARTAK%26AND%26znacaj%2bodrzavanja%2b%2boralne%2bhigijene&page=0&sort=1&stype=1&backurl=%2fSearchResults.aspx%3fquery%3dARTAK%2526AND%2526znacaj%252bodrzavanja%252b%252boralne%252bhigijene%26page%3d0%26sort%3d1%26stype%3d1

[6] Ljušković Lj, Tošovice V, Hasanagić S, Janković S. Maintenance of oral hygiene in patients with fixed orthodontic appliances. Stomatološki informator. 2011; 11 (27) 11: fifth [Cited 2016 Mar 30]. Available from: http://dlv.org.rs/wp-content/uploads/2015/05/StomatoloskiInformator27comp.pdf

[7] Računica J, Ivetic V, Naumović N, Durić M. TheeEffect of smoking on resorption of alveolar bone. Serbian Dental Juornal. 2008; 55 (2): 107-14. [Cited 2016 Feb 10]. Available from: http://web.b.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=00391743&AN=33059850&h=MzHocoKW7UVXYEO%2bqNg3w8h19p92lz%2fe0FaFBrgEaAFsOw7%2baiNgKpyJBkADJw4t38vhZ1CuNjwDhvcNkK9sAQ%3d%3d&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d00391743%26AN%3d33059850

[8] Grgic O, Blagojević D. The influence of diet on oral health. Dental informant. 2012 [cited 2016 Feb 6]. Available from: http://scindeks-clanci.ceon.rs/data/pdf/1451-3439/2012/1451-34391231016G.pdf#search="oralno

[9] Peres MA, Sheiham A, Liu P, Demarco FF, Silva AE, Assuncao MC, Menezes AM, Barros FC, Peres KG. Sugar consumption and change in dental caries from chilhood to adolescence. J Dent Res. 2016; [Cited 2016 Feb 6]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26758380

[10] Verzak Z, Burazin A, Černi I. Fluoride and caries. Medix. 2007; 71: 155-6. [Cited 2016 Feb 6]. Available from: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=91154

[11] von der Fehr FR. Caries prevalence in the Nordic countries. International Dental Journal. 1994; 44 (4): 371 to 8. [Cited 2016 Feb 2]. Available from: http://europepmc.org/abstract/med/7814104

[12] Usaman S, Bhat S, Sargodha S. Oral helath knowledge and behavior of clinical medical, dental and paramedical students in Mangalore. J Comm Dent Oral Health. 2007; 1 (3): 46-8. [Cited 2016 Feb 2]. Available from: http://johcd.org/pdf/JOHCDOral%20Health%20Knowledge%20and%20Behavior(1).pdf

[13] Šimat S, Mostarčić K, Matijević J, Simeon P, Grget K.R, Krmek J.S. Comparison of oral status of the fourth year students of various colleges at the University of Zagreb. Acta Stomatologica. 2011; 45 (3): 177 to 83. [Cited 2016 Feb 2]. Available from: http://hrcak.srce.hr/71728

[14] Nadeem M, Sidra S, Ahmed S, Khaliq R, Mirza H. Evaluation of dental health education and dental status among dental students at Liaquat College of Medicine and Dentistry. International Journal of dental clinics. 2011; 3 (3): 11-3. [Cited 2016 Jan 3]. Available from: http://intjdc.org/index.php/intjdc/article/view/3.3.4/pdf

[15] Duričkivić M, Ivanović M. The state of oral health in children at age of 12 in Montenegro. Vojnosanitetski Pregled. 2011; 68 (7): 550-5. [Cited 2016 Feb 2]. Available from: http://scindeks-clanci.ceon.rs/data/pdf/0042-8450/2011/0042-84501107550D.pdf

[16] Akar C.A. Özmutaf N. M, Ozgur Z. An assessment of self-reportes oral health behavior of non-dental school students in Turkey. Acta Stomatologica Croatica. 2009; 43 (1): 13- 23. [Cited 2016 Mar 9]. Available from: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=54499

[17] Jaramillo JA1, Jaramillo F, Kador I, Masuoka D, Tong L, Ahn C, Komabayashi T. A comparative study of oral health attitudes and behavior using the Hiroshima University- Dental Behavioral Inventory (HU-DBI) between dental and civil engineering students in Colombia. J Oral Sci. 2013;55(1):23-8. [Cited 2016 Jun 3]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4090926/pdf/nihms596748.pdf

[18] Davidovic B, Ivanović M, Jankovic S, Lečić J. Knoweledge,attitudes and behavior of children in relation to oral health. Vojnosanitetski Pregled. 2014; 71 (10): 949-56. [Cited 2016 Mar 9]. Available from: http://scindeks.ceon.rs/article.aspx?query=ARTAK%26AND%26oralno%2bzdravlje%2b&page=9&sort=1&stype=1&backurl=%2fSearchResults.aspx%3fquery%3dARTAK%2526AND%2526oralno%252bzdravlje%252b%26page%3d0%26sort%3d1%26stype%3d1

[19] Chrysanthakopoulos N.A. Self-reported oral health attitude and behaviuor of Greek medical students. Acta Stomatologica Croatica. 2012; 46 (2): 126 to 35. [Cited 2016 Mar 12]. Available from: http://hrcak.srce.hr/84246

[20] Sato M1, Camino J, Oyakawa HR, Rodriguez L, Tong L, Ahn C, Bird WF, Komabayashi T. Effect of dental education on Peruvian dental students' oral health-related attitudes and behavior. J Dent Educ. 2013;77(9):1179-84. [Cited 2016 Jun 4]. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4090930/pdf/nihms596747.pdf

[21] Cvetković J, Nenadović M, Stojanović Tasic M, Milošević N. tobacco usage among the student population in North Kosovo. Biomedical research. 2015; 6 (1): 46-51. [Cited 2016 Mar 10].

[22] Bokan D, Bokan D, Rakić D, Budakov N. Prevalence of tobacco smoking among students of the University of Novi Sad. South Eastern Europe Health Sciences Journal. 2012; 2 (2). [Cited 2016 Mar 10]. Available from: http://unvi.edu.ba/SEEHSJ/volume_2_no2/Dalibor%20Bokan1%20SEEHSJ%20novembar%202012.pdf

[23] Cvetković A, Vulović M, Ivanović M. Correlation between dental health status and environmental factors: nutrition, oral hygiene and saliva in children. Serbian Dental Journal. 2006; 53: 217-28. [Cited 2016 Mar 15]. Available from: http://www.doiserbia.nb.rs/img/doi/0039-1743/2006/0039-17430604217C.pdf