FINDING OVERLAPPING TERMS IN MEDICAL AND HEALTH CARE CURRICULUM USING TEXT MINING METHODS: REHABILITATION REPRESENTATION – A PROOF OF CONCEPT

Matěj Karolyi[1*], Martin Komenda[1], Radka Janoušová[2], Martin Víta[3], Daniel Schwarz[1]

[1] Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic

[2] Faculty of Medicine, Masaryk University, Brno, Czech Republic

[3] Faculty of Informatics, Masaryk University, Brno, Czech Republic

** Corresponding author: karolyi@iba.muni.cz*

ARTICLE HISTORY
Received 27 December 2017
Accepted 11 January 2017
Available online 17 January 2017

KEYWORDS
natural language processing;
curriculum;
education;
medical;
rehabilitation

ABSTRACT —***Background:*** *Various institutions dealing with higher medical and healthcare education have different methods of organising their study programmes, which typically involve hundreds of theoretically and clinically focused courses. The importance of a well-balanced curriculum is indisputable – the society needs qualified doctors because people's health is necessary for the functioning and development of the entire society.*

Objectives: *In this paper, we introduce our innovative approach to identify overlaps among medical or healthcare disciplines using term similarity. A close attention is focused on the discipline of Rehabilitation and Physical Medicine and its role in the General Medicine study field in the Faculty of Medicine at Masaryk University.*

Methods: *Data and text mining techniques were used in practice, in accordance with a time-tested methodological background, which systematically covers all fundamental steps to discover and to extract knowledge from data repositories. In order to extract term similarities from a medical curriculum dataset, the CRISP-DM reference model was chosen as a well-documented practical guideline.*

Results: *The achieved results clearly demonstrate overlapping areas among the defined disciplines in the explored curriculum. The resulting comprehensive analytical report presents the term occurrence in a set of figures and tables, which were thoroughly evaluated by experts familiar with the curriculum design process.*

Conclusions: *In this case study, we have proposed an innovative method for identifying overlaps of terms occurring in medical and healthcare disciplines when compared to the discipline of Rehabilitation and Physical Medicine. The first results are promising in the sense of face validity. We believe that this approach can be used similarly to gain an objective overview of the entire curriculum.*

INTRODUCTION

Correctly compiled and balanced curricula are an essential prerequisite for undergraduate medical and healthcare education [1]. Various institutions dealing with this kind of education have different methods of organising their study programmes, which typically involve hundreds of theoretically and clinically focused courses. Their curricula – in the form of compulsory, compulsorily optional, and optional courses – are available to students and teachers in various local online environments. These platforms, typically learning management systems (LMSs) and curriculum management systems (CurrMSs), can provide a very comprehensive description of the entire study programme structure including a detailed curriculum content.

Medicine is one of the widest branches of knowledge. The progress in medical research (including new findings, therapeutic procedures, laboratory and imaging methods etc.) is very fast [2–4]. That is why the compilation of the General Medicine (GM) curriculum is a very complex problem, which is really difficult to solve. However, the importance of a well-balanced curriculum is indisputable – the society needs qualified physicians who will look after people's health. Moreover, most physicians make every effort to treat patients according to their best knowledge and conscience. A well-balanced curriculum must guarantee that each medical student gets as much knowledge in a specific branch of medicine as it is needed for him/her to be able to start working in the corresponding department after his/her graduation and, at the same time, he/she should get enough knowledge of all other branches of medicine in order to be able to effectively cooperate with his/her colleagues and to make appropriate decisions on diagnostic and therapeutic procedures.

The last decade has seen an increasingly growing trend towards the integration of automated natural language processing (NLP) techniques for obtaining relevant information from large amounts of data from the medical domain [5–7]. As regards the domain of curriculum innovation and its complexity, these methods could explore information-rich relations between the individual curriculum parts and contribute to a better transparency of a global overview. In this paper, we introduce our innovative approach to identify medical and healthcare disciplines overlaps using term similarity. A close attention is focused on one particular discipline, namely the Rehabilitation and Physical Medicine (RPM), and its role in the GM study field. We have investigated an effective model for a computer-assisted discipline comparison based on keywords representation. In our particular case, we have deeply analysed selected curriculum text-based descriptive parameters from the OPTIMED CurrMS [8]. OPTIMED has been developed and implemented in the Faculty of Medicine at Masaryk University (FM MU) in Brno, Czech Republic, and nowadays provides an essential support during curriculum planning, creation, management, and optimisation. It already covers a huge amount of data: 1,347 learning units and 6,974 learning outcomes, i.e. more than 2,500 standard pages of text.

The role of rehabilitation in the curriculum of the General Medicine study programme

General Medicine (GM), the largest study field at FM MU, does not involve any separate RPM course. RPM is taught as part of two courses, namely Internal Medicine – Block 5, and Orthopaedics and Rehabilitation – Practice [9]. Apart from these two courses, a computer-assisted search for the representation of this discipline has to be done in all other courses of GM; in this way, we can find out whether the entire curriculum contains essential concepts of this medical discipline or not.

In general, RPM is an interdisciplinary branch of medicine closely linked to all other medical disciplines. Professor Pavel Kolář [10] defines the rehabilitation as follows: “Treatment rehabilitation is an integral part of healthcare and includes a complex of rehabilitative, diagnostic, therapeutic and organisational actions directed toward reaching an individual’s maximum functional potential and establishing conditions for its achievement. Treatment rehabilitation can be provided in the form of inpatient care, outpatient care and specialised care in treatment institutions, including balneological centres. It should be initiated in all areas of clinical specialties, including intensive care units (ICU) during the period of acute inpatient medical care.”

The postgraduate education of physicians encompasses a lot of interconnections between rehabilitation topics and other medical disciplines, too. In the Czech Republic, a graduate student must complete the so-called specialised education of physicians, involving a two-year common trunk followed by a basic branch (the final specialisation) to be allowed to work individually [11]. The majority of these branches are preceded by only one or two common trunks (for example, the internal medicine trunk or the paediatric medicine trunk precede the infectious medicine branch; the internal medicine trunk is a prerequisite of the angiology branch, etc.). However, the choice of the common trunk preceding the RPM branch is wider: a graduate student can choose one of five different common trunks to continue with the RPM branch, specifically the internal, surgical, paediatric, orthopaedic or neurological trunk.

The above-mentioned definition of rehabilitation and the postgraduate education description, as well as the experience from the clinical stage of medical education, imply that the knowledge of this discipline is really important for medical students and graduates. If a physician – a specialist in any medical discipline – has at least basic knowledge of rehabilitation, he/she will be able to apply it in his/her practice or contact a colleague specialised in rehabilitation at the right time and send a patient with mobility problems to the right place.

The global objective of this study was to find out how to measure the content similarities – based on keywords occurrence – among medical and healthcare disciplines from OPTIMED and RPM.

METHODS

Using NLP techniques in practice should be based on a time-tested methodological background, which systematically covers all fundamental steps to discover and to extract knowledge from collected datasets. In general, standards establishment in the data-mining sphere always has to be taken into account. For the purpose of finding term similarities in medical curriculum data, the CRoss-Industry Standard Process for Data Mining (CRISP-DM) reference model was chosen as a well-documented practical guideline [12]. CRISP-DM divides the exploratory process into six major phases, as described below.

Business understanding

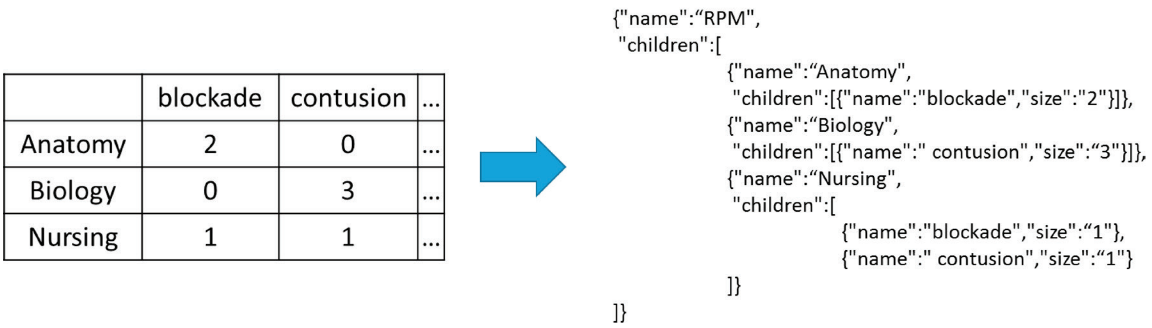
First of all, domain understanding and formulation of the global objectives are required. Here, we used the dataset collected within OPTIMED CurrMS, which provides an essential support during institutional decision-making activities relating to the curriculum creation, management, and optimisation [1]. In this system, a curriculum is made of fundamental building blocks (such as study field, discipline, course, learning unit, and learning outcome) supplemented with descriptive attributes (such as teaching range and type, annotation, keywords, and information resources). OPTIMED CurrMS provides data about all 42 medical and healthcare disciplines except RPM. Based on these data, we were able to merge the selected attributes and to create textual representations for a further NLP discipline comparison. As RPM does not constitute a separate discipline or a course at FM MU, no data describing RPM were available. In order to have an adequate RPM keyword representation, a new RPM description standing out of OPTIMED was artificially defined by three RPM experts. It consisted of 110 fundamental terms divided into the three classes (Table 1).

Data understanding

The concrete dataset that we used to measure content similarities with RPM was extracted from the above-mentioned 42 disciplines in the OPTIMED system. The main content of these disciplines is logically stored in learning units and learning outcomes. The curriculum of FM MU is currently described by 1,347 learning units and by 6,974 learning outcomes. In general, we are able to export all of these data using predefined or customisable exporting tools integrated within the OPTIMED platform as a text (free form) or as HTML tables (structured form). For the purpose of this study, we processed a comma-separated textual dump of learning units (approximately 6,149 kB of data) and learning outcomes (approximately 3,574 kB of data). The key attributes of each learning unit involved their importance, annotation (long text attributes), Medical Subject Headings (MeSH) standardized

TABLE 1. Overview of 110 terms in three categories

Basic terminology	Analgesy, balneotherapy, biomechanical analysis, coding theory, comprehensive rehabilitation, convalescence, counternutation, deep stabilization system, developmental kinesiology, dispersion efect, elasticity, endorphin theory, ergotherapy, gate theory, goniometry, interstitial neuron, joint play, motoricity, muscle testing, muscle tone, myorelaxation, myostimulation, neuromuscular, neuromuscular coordination, nutation, palpation, physiatry, physiotherapy, post-isometric relaxation, psychosomatic medicine, pulmonary rehabilitation, reciprocal innervation, reconditioning, rehabilitation, rehabilitation medicine, rigidity, sensorimotor, stereotypic movement, thixotropy, trace element, vertebrogenic, vocational rehabilitation.
Pathology	Algodystrophy, arthrosis, avascular/aseptic necrosis, blockade, cerebral palsy, cervicalgia, complex regional pain syndrome (CRPS), contusion, conversion disorder, distortion, enthesopathy, functional disorder, hemiparesis, hyperalgesic area, hypermobility, chaining, chondromalacia patella, instability, key area, ligament laxity, lumbago, muscle dysbalance, myogelosis, pain, paraparesis, rupture, scoliosis, slipped disc, stress fracture, tendinitis, tenosynovitis, trigger point, upper crossed syndrome, vertebrocardiac syndrome, vertebrogenic algic syndrome (VAS), vertebrovisceral syndrome.
Therapy	Bobath concept, cardio workout, catelectrotonus, closed chain exercise, cryotherapy, dynamic neuromuscular stabilization (DNS), electrogymnastics, electrotherapy, HIL therapy, interfering current, laser therapy, long-term rehab program, lymphatic drainage, magnetotherapy, manipulation, massage, mechanotherapy, mobilization, Mojzis methods, orthosis, peloids, physical therapy, proprioceptive neuromuscular facilitation, prosthetics, reflex locomotion, reflexology, shockwave, soft tissue techniques, therapeutic exercise, thermotherapy, Vojta therapy, water therapy.



	blockade	contusion	...
Anatomy	2	0	...
Biology	0	3	...
Nursing	1	1	...

```
{"name":"RPM",
 "children":[
              {"name":"Anatomy",
               "children":[{"name":"blockade","size":"2"}]},
              {"name":"Biology",
               "children":[{"name":" contusion","size":"3"}]},
              {"name":"Nursing",
               "children":[
                            {"name":"blockade","size":"1"},
                            {"name":" contusion","size":"1"}
              ]}
]}
```

FIGURE 1. Dataset transformation from CSV to JSON

keywords (one to five words), significant terms (set of words or phrases in the tree structure) and a list of linked learning outcomes according to the Bloom's taxonomy [13]. These dumps also included pieces of information irrelevant for this experiment, and these had to be removed (this process is described in the data preparation section). Furthermore, we also processed data on RPM which were not included in the OPTIMED CurrMS. In contrast to the description of the other 42 disciplines, this dataset consisted only of 110 terms divided into three subsections.

Data preparation

Before the core measurement of content similarities, the data were preprocessed and cleaned. This was done by several subsequent techniques. The first one involved a supervised filtering of attributes via the OPTIMED custom export of learning units. In general, the user is able to specify which attributes of learning units (e.g. annotation, title, MeSH keywords, significant terms etc.) should be included to the final comma-separated values (CSV) file. This custom report is available for learning outcomes, and therefore they need to be filtered manually in third-party tools, e.g. Microsoft Excel. In the next step of data preparation, special characters were removed: in particular, HTML tags used for additional formatting were eliminated. In the last step of data preparation, the so-called stop words were removed. Data were cleaned from the verbal ballast, i.e. words such as conjunctions, prepositions and manually selected words from a predefined list. After this preprocessing step, the dataset was ready to be used in a further analysis. Further processing was performed using the TM package in R (https://cran.r-project.org/web/packages/tm/index.html) in order to obtain a document-term matrix. This procedure consisted of: (i) lowercasing, (ii) replacing multiword keyword expressions by "artificial underlined words", (iii) removing punctuation, (iv) removing numbers, (v) removing stop-words (as provided as stop-words within tm package), and (vi) removing whitespaces.

Modelling

A document-term matrix, which uses term frequency weighting, covers only words (tokens) of the length of at least 3 characters within the learning units; that means, words containing only one or two characters were ignored. From the entire document-term matrix, only a selection of words was used for further consideration: we therefore worked with a reduced matrix, namely with columns (words) that were both involved in the rehab-terms.txt (110 keywords of RPM defined by experts) and in at least one plaintext file that represented a certain discipline. Words from rehab-terms.txt that did not appear in any of the documents were stored in a special list (missing-words.txt).

The document showing overlaps between RPM and other disciplines was named as reducedDTM.txt and its content had the form of a matrix where the names of columns were keywords defined by experts with at least one occurrence in other disciplines; these disciplines were in rows of the matrix. For the purpose of visualisation, we transformed the data structure to

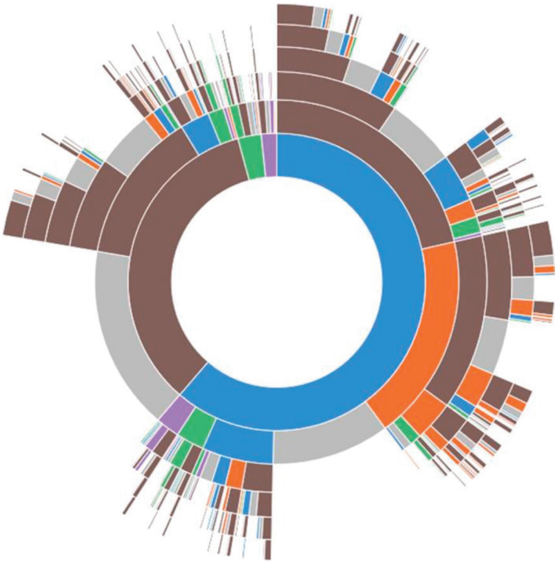


FIGURE 2. A general example of a sunburst graph (https://bl.ocks.org/)

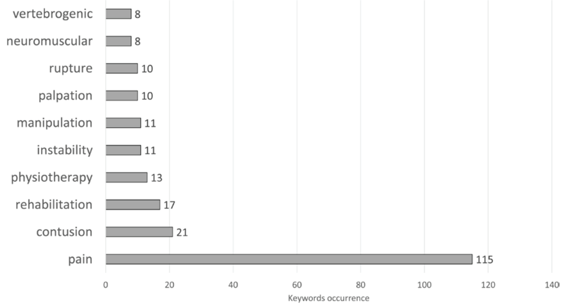


FIGURE 3. RPM keyword occurrence

TABLE 2. Overlaps of medical disciplines with RPM

Non-overlapping disciplines	Anatomy, Basic Medical Terminology, Health Care and Policy, Histology and Embryology, Medical Chemistry, Medical Microbiology, Surgery.
Overlapping disciplines	Anaesthesiology and Treatment of Pain, Biology, Biophysics, Clinical Examination in Internal Medicine, Clinical Oncology, Communication and Self-Experience, Community Medicine, Dermatovenerology, Diagnostic Imaging Methods, Epidemiology of Infectious Diseases , Family Medicine and Geriatrics, First Aid, Forensic Medicine, Gynaecology and Obstetrics, Immunology, Infectious Diseases, Intensive Care Medicine, Internal Medicine, Medical Ethics, Medical Psychology, Neurology, Neuroscience, Nursing, Ophthalmology, Orthopaedics, Otorhinolaryngology, Pathological Physiology, Pathology, Paediatrics, Pharmacology, Physiology, Preventive Medicine, Psychiatry, Stomatology.

a more suitable format, namely the JavaScript Object Notation (JSON). The input and output examples of algorithm for the above-mentioned modification are shown in Figure 1.

Evaluation

At this phase, RPM experts not involved in the OPTIMED project critically assessed the achieved results in terms of medical curriculum keywords overlaps or absence. Based on the summarising report in the form of data tables and interactive graphs, experts commented on the keywords occurrence among medical and healthcare disciplines from OPTIMED and RPM. After this evaluation process, the presented solution was ready to deploy.

Deployment

The final visualisation was divided into three main parts and it can be seen online in the OPTIMED CurrMS (http://opti.med.muni.cz/en/reporting/web/analyticke-reporty/rehabilitace-ve-vyuce). The first part is a list of the RPM keywords missing in other disciplines (missing-words.txt). The second part shows the list of disciplines having no intersection with the field of Rehabilitation and Physical Medicine. This means that not a single word from the above-mentioned 110 terms is present in the plaintext file of that discipline. The list of missing disciplines is a by-product derived from the last step of data preparation. It illustrates the intersections of RPM keywords with other 42 disciplines. We used the sunburst chart (see Figure 2) to display hierarchical data, more specifically the three-layered tree. RPM corresponds to the root, the other 42 disciplines are in the second layer, and terms are located in the third layer. The size of each sector depends on the number of occurrences of each term in the respective discipline.

RESULTS

Our results show how the disciplines from OPTIMED CurrMS are connected to RPM. Table 2 represents two sets of disciplines: (i) disciplines not overlapping with RPM terms, (ii) disciplines overlapping with RPM

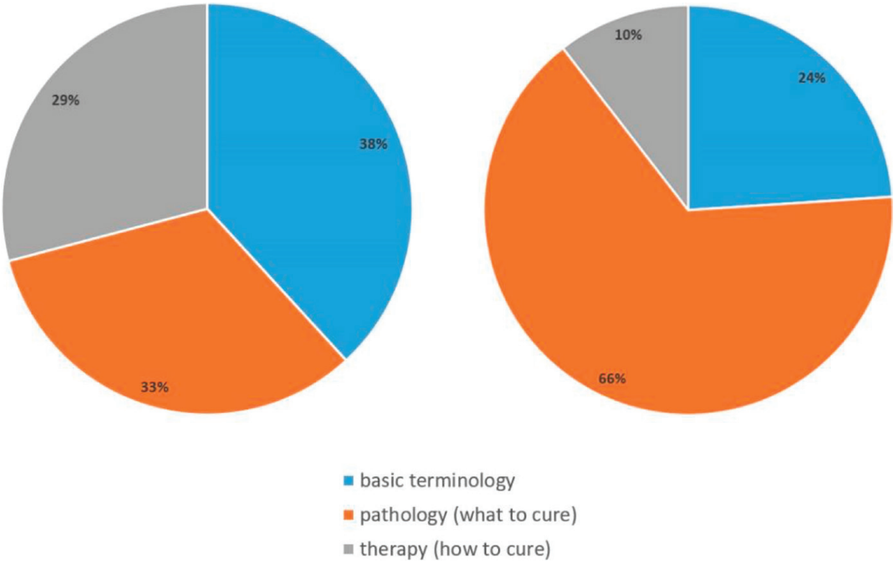


FIGURE 4. RPM classes in accordance with total keyword occurrence

terms, containing at least one occurrence of any fundamental RPM keyword. In our case, seven out of 42 disciplines had no overlap with RPM and were therefore excluded from further analysis.

The most frequently occurring word was "pain" (see Figure 3), which was found 115 times across all disciplines' files. The frequency of other words is significantly lower.

Figure 4 shows the coverage of three RPM subsections of terms. The left pie chart shows the relative sizes of input sets, whereas the right pie chart shows the relative sizes of sets in which keywords overlap with other 42 medical disciplines. The result is unbalanced, considering the fact that the basic terminology provides the biggest set of keywords, and its portion is only 24 per cent.

DISCUSSION

RPM experts not involved in the OPTIMED project critically commented two principal points resulting from our study. Firstly, the methodology evaluation: the methods used in our study (and the achieved results as well) reflect on the actual conditions of RPM education (or any other discipline which was not originally included in OPTIMED). The model is suitable to provide comparisons among disciplines which are included in the entire GM curriculum. Secondly, the RPM education overview: the RPM education does not play a sufficient role in the GM curriculum at FM MU according to results of this study, considering its impact to the medical knowledge and the professionalism of future physicians.

Methodology evaluation

RPM experts agreed that this model is very useful for a general overview about keywords representation in each discipline. The model brings a unique possibility to compare taught medical and healthcare disciplines in the levels of learning units and learning outcomes; this can be very helpful in any optimisation efforts related to curriculum management. Our model is therefore a functional instrument for the achievement of goals for which OPTIMED was developed.

Our model has several limitations. Firstly, the English mutation of OPTIMED was used to compare terms across disciplines due to the difficulty of automatic natural language data processing in the Czech language. But the conformity of using terminology in both languages can be questionable because of necessity of translation Czech medical terms to English ones. Moreover, the practical use of OPTIMED has shown slight discrepancies between terms representing individual disciplines in Czech and English language mutations. The OPTIMED curriculum browser in the Czech language found several of 84 fundamental RPM terms among learning units and learning outcomes despite the fact that these terms could not be found across all disciplines' files in our study. The second limitation arises in cases where there are too many universal terms with a vague meaning (e.g. "pain, contusion, instability, palpation" etc.). These terms are included in descriptions of multiple disciplines, but they represent very universal and common symptoms or processes in medicine. Consequently, their representation in the curriculum cannot guarantee the education of RPM specifically. And thirdly, one expert pointed out that just a simple occurrence of a medical term in a learning unit or a learning outcome cannot assure that the term is interpreted and understood correctly by students.

RPM education overview

Our results demonstrate that the described curriculum at FM MU does not contain a large part of RPM keywords. The essential terms "rehabilitation" and "physiotherapy" have been found only 17 times and 13 times respectively in the whole dataset. That is a very low frequency considering the fact that the RPM is closely linked to all other medical disciplines. The missing RPM basic terms describe treatment methods which are used in many patients in everyday clinical practice. For example, GM students at FM MU do not learn anything about the "Vojta therapy", which is a globally recognised and widely used therapeutic method developed by a Czech neurologist, Professor Vaclav Vojta. Other terms, namely "cryotherapy", "magnetotherapy" and "electrotherapy" have been found only 2 times in the dataset, but such a low frequency does not correlate with the widespread use of these therapeutic methods.

The study brings interesting results from the analysis of keywords' frequency in various disciplines. We were glad to see a relatively numerous representation of RPM terms in the Orthopaedics discipline. We were rather surprised at the presence of First Aid and Forensic Medicine among six disciplines with the largest numbers of RPM keywords. On the contrary, the results show a poor representation of RPM basic terms in many disciplines such as Surgery, Paediatrics and Preventive Medicine, although these medical disciplines are closely linked to RPM in clinical practice. There are various explanations for these results. For example, descriptions of different disciplines might not contain the same terms at the same level of detail. In our study, RPM terms were divided into three classes. Keywords of the Therapy group covered only 10% of all terms of the entire curriculum. One of our experts pointed it out with an explanation that it might reflect the fact that the GM education puts more emphasis on the pharmacological or surgical treatment despite their costs and side effects.

CONCLUSION

In this case study, we have designed, performed and evaluated an innovative method for identifying overlaps of terms between a whole medical curriculum and the specific discipline of Rehabilitation and Physical Medicine. The achieved pilot results seem to be promising in terms of face validity. We believe that the presented method can provide an objective curriculum overview described by individual medical and healthcare disciplines. We concluded that the RPM education is not represented sufficiently in the GM curriculum at FM MU. Its coverage in this specific curriculum does not correspond to the importance of the RPM discipline in the domain of clinical medicine. RPM experts agreed that results of our study are very consistent with the reality of current conditions in clinical medicine. Based on our results, experts concluded that the absence of a separate RPM course is not well substituted by other GM courses. In general, the global reasonability and usability of this model in practice has a big potential. As indicated above, the model is a functional instrument for the achievement of goals for which OPTIMED was developed. Our case study shows an innovative way of measuring medical curriculum content similarities using the keyword occurrence.

Matěj Karolyi

ACKNOWLEDGEMENTS

The authors would like to thank to RPM experts, namely Jana Martinková, MD, Radana Poděbradská, PhD, and Radek Brauner, MD, who actively participated on the evaluation part of this study, and provided an objective summary with relevant comments.

REFERENCES

[1] Komenda M. Towards a Framework for Medical Curriculum Mapping. PhD thesis, Masaryk University, Faculty of Informatics, 2016.

[2] Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform 2016. DOI: 10.1093/bib/bbw065

[3] Vercher-Conejero JL, Pelegrí-Martinez L, Lopez-Aznar D, del P. Cózar-Santiago M. Positron emission tomography in breast cancer. Diagnostics 2015; 5(1): 61–83.

[4] Calvet D, Mas JL. Recent advances in carotid angioplasty and stenting. Int J Stroke 2016; 11(1): 19–27.

[5] Hersh WR, Gorman PN, Biagioli FE, Mohan V, Gold JA, Mejicano GC. Beyond information retrieval and electronic health record use: competencies in clinical informatics for medical education. Adv Med Educ Pract 2014; 5: 205–212.

[6] Dajun T, Xu Z. Analysis on curriculum of information retrieval of library and information science in America. Libr Inf Serv 2014; 15: 17.

[7] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003; 36(6): 462–477.

[8] Komenda M, Schwarz D, Vaitsis C, Zary N, Štěrba J, Dušek L. OPTIMED platform: curriculum harmonisation system for medical and healthcare education. Stud Health Technol Inform 2015; 210: 511–515.

[9] Táborská E, Neckařová M. Faculty Calendar 2016/2017. [Online]. Faculty of Medicine, Masaryk University, 2016. [cit. 2016-Oct-07]. Available at WWW: <http://www.med.muni.cz/index.php?id=10>.

[10] Kolář P et al. Clinical Rehabilitation. Alena Kobesová, 2014. ISBN 978-80-905438-1-2.

[11] IPVZ. Jak získat specializovanou způsobilost. [Online]. [cit. 2016-06-24]. Available at WWW: <https://www.ipvz.cz/lekari-zubni-lekari-farmaceuti/ziskavani-specializace/jak-ziskat-specializovanou-zpusobilost>.

[12] Chapman P et al. CRISP-DM 1.0 Step-by-step data mining guide. SPSS 2000.

[13] Huitt W. Bloom et al.'s taxonomy of the cognitive domain. [Online]. Educational Psychology Interactive. Valdosta State University, 2011. [cit. 2016-10-29]. Available at WWW: <http://www.edpsycinteractive.org/topics/cognition/bloom.html>.