A peer-reviewed version of this preprint was published in PeerJ on 25 November 2014.

View the peer-reviewed version (peerj.com/articles/683), which is the preferred citable publication unless you specifically need to cite this preprint.

Vaitsis C, Nilsson G, Zary N. 2014. Visual analytics in healthcare education: exploring novel ways to analyze and represent big data in undergraduate medical education. PeerJ 2:e683
https://doi.org/10.7717/peerj.683

Visual Analytics in healthcare education: Exploring novel ways to analyze and represent big data in undergraduate medical education

Introduction

Big data in undergraduate medical education that consist the medical curriculum are beyond human abilities to be perceived and analyzed. The medical curriculum is the main tool used by teachers and directors to plan, design and deliver teaching activities, assessment methods and student evaluation in medical education in a continuous effort to improve it. It remains unexploited mainly for medical education improvement purposes. The emerging research field of Visual Analytics has the advantage to combine data analysis and manipulation techniques, information and knowledge representation, and human cognitive strength to perceive and recognize visual patterns. Nevertheless, there is lack of findings reporting use and benefits of Visual Analytics in medical education.

Methods

We analyzed data from the medical curriculum of an undergraduate medical program concerning teaching activities, assessment methods and results and learning outcomes in order to explore Visual Analytics as a tool for finding ways of representing big data from undergraduate medical education for improvement purposes. We used Cytoscape to build networks of the identified aspects and visualize them.

Results

The analysis and visualization of the identified aspects resulted in building an abstract model of the examined data from the curriculum presented in three different variants; (i) learning

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.456v1 | CC-BY 4.0 Open Access | received: 11 Aug 2014, published: 11 Aug

outcomes and teaching methods, (ii) examination and learning outcomes and (iii) teaching methods, learning outcomes, examination results and gap analysis

Discussion

This study identified aspects of medical curriculum. The implementation of VA revealed three novel ways of representing big data from undergraduate medical education. It seems to be a useful tool to explore such data and may have future implications on healthcare education. It also opens a new direction in medical informatics research.

Christos Vaitsis[1], Gunnar Nilsson[2] and Nabil Zary[1]

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

[2]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden

Corresponding author:

Christos Vaitsis,

Department of Learning, Informatics, Management and Ethics,

Karolinska Institutet,

17177 Stockholm, Sweden,

Tel: +46 (0) 73 71 21 547

Email: christos.vaitsis@ki.se

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.456v1 | CC-BY 4.0 Open Access | received: 11 Aug 2014, published: 11 Aug

Introduction

Medical Education

Continuous efforts to improve medical education today are currently driven from the need to create competent health professionals able to meet healthcare demands. One approach has been to react to observed deficiencies in healthcare that were linked to unsatisfactory required competencies (Frenk et al., 2011). An example was given by Cho et al. who highlighted the difficulties the physicians had to keep pace with the growing medical literature, and therefore proposed to include a journal club in undergraduate medical education to acquire at an early stage the ability to critically review scientific literature, a skill needed as a future physician using evidence-based medicine (Cho et al., 2011). While reviewing the literature, we found limited studies reporting on improvement of medical education based on educational data such as assessment results and evaluation data (Gathright et al., 2009). Corrin and Olmos analyzed data from medical education, and more particular reflective records from clinical experiences through an online clinical log system and showed that it could help medical faculties to enhance the alignment between medical students' clinical experiences and the taught curriculum (Corrin & Olmos, 2010). In another study, non - previously perceived discrepancies between taught and the assessed curriculum in medical program were revealed using a web-based learning objectives database (Hege et al., 2010).

Big Data

Big Data is broadly defined today as the existence or emergence of datasets with such a magnitude that is beyond recently used tools (mainly databases) abilities to warehouse, manipulate and analyze. Different sectors like the public, commercial and social, receive and produce every day, hour and minute vast amounts of data from different sources and in different forms. The large size of these data in terabytes or even petabytes exceeds the hardware or human abilities to easily process them and therefore they are characterized as Big Data. Nevertheless, this term is arbitrarily given to large sized data, and it can vary from sector to sector and more specifically between services within a sector (Manyika et al., 2011).

The size of data is only one characteristic that can easily confer the Big Data term to them. Other characteristics that define Big Data today except size (referred as volume) are variety and velocity. Variety refers to the different types the data can be found and the different sources they can be collected from, in structured and unstructured forms. Velocity refers to the speed the data

are produced but also to the time they are processed, in real-time or occasionally (Eaton et al., 2012).

Big Data from Higher Education

Higher education is one of the domains where data frequently collected from students' usage and interaction, course information and other academic data like administration and curricula, are in such size and type that special techniques must be applied to discover new knowledge. (Romero & Ventura, 2007)

It is reported how within the context of higher education, Big Data have the potential to enable the development of insights ''*regarding student performance and learning approaches*'' and gives examples of areas in big educational data - like student's actual performance according to taught curriculum - that can be positively affected. (West, 2012)

Additionally, Big Data and analytics in higher education have recently been seen as a great potential to promote actions. These actions concern ''*administrative decision-making and organizational resource allocation*'', early identify students at risk and intervene to prevent them from failing, develop more effective instructional techniques and transform the traditional view of the curriculum into a network of relations, using educational data collected regularly from learning management systems, social networks, learning activities and the curriculum (Siemens & Long, 2011).

Between identified areas in which Big Data and analytics can be used for investigation and improvement in higher education is the curriculum and its contents, as part of educational data (Picciano, 2012).

Complexity of Higher Medical Education

The medical curriculum is inherently complex from the multi-aspect nature of medical education (Maojo et al., 2002; Mennin, 2010). The rapidly changing world of healthcare imposes the existence of a flexible healthcare education system and consequently of a flexible medical curriculum that can be analyzed and used as a base to support and inform changes and improvements in healthcare education. Aligned with this philosophy and additional anticipating to provide a way to reduce the complexity of a medical curriculum and transform it into an understandable and interoperable tool to facilitate in developing and qualitative improving ''*health professions education curricula*'', the Medbiquitous organization has ''*developed and promote technology standards for the health professions that advance lifelong learning,*

continuous improvement, and better patient outcomes'' (http://medbiq.org/). These technology standards use terminology in structured Extensible Markup Language (XML) format that describe the different parts of a medical curriculum.

To make this study interoperable in terms of research or even benchmarking purposes, this study will provide pairings of terminology used in the examined medical curriculum to Medbiquitous terminology.

Curriculum data currently used in education in the undergraduate medical program in Sweden and are available to medical program/courses directors, teachers and developers (from now defined as stakeholders) exist in different places and in different forms and sources. Those are:

- The ones defined from the Swedish Higher Education Authority (higher education board) (http://english.uk-ambetet.se) and describe the intended learning outcomes (sixteen LO1-LO16 in Appendix S4) of the medical program in national level
- Those in medical program's and each course's webpage along with the respective syllabus of the course where all learning activities, assessments and learning outcomes are described and
- In a description of the whole medical program at some universities in an educational database (https://internwebben.ki.se/sv/Selma).

In addition, another source that the same curriculum data can be found and concern the whole healthcare education from the highest level of the higher education board to program and courses are those collected by medical programs prior to their external evaluation. Apart from the primarily reason for external evaluation, these data were created in an effort to transform Big Data from the medical curriculum to an auxiliary instrument to support education development and improvement and also to create a comprehensible overview of the courses and the whole medical program. Nevertheless, the form of data as text and numbers in numerous worksheets and the level of complexity are comprehensible to a certain extent to those who created the data. They are not yet available to different stakeholders in healthcare education who only have access to curriculum data in different forms and sources as described before.

A possible use of Big Data in medical education is to:

- Identify connections and relations between different entities in all levels
- Determine the role of each entity in the lowest level of a course but also to the overall picture of the medical program
- Perceive and analyze the curriculum in terms of identifying if knowledge, skills and attitude are constructed through the alignment of teaching methods and assessment

towards the learning outcomes which is called constructive alignment and is defined from (Biggs & Tang, 2007), between different entities in the medical curriculum

- Perform gap analysis (Gannod, Gannod, & Henderson, 2005; Ritko & Odlum, 2013) in terms of comparing different states an entity can be found to identify possible discrepancies and ensure curriculum's alignment between intended and actual curriculum, in-between all different levels of the curriculum but also the curriculum's structure towards the defined learning outcomes from Swedish Higher Education Authority

Performing these actions on medical curriculum is similar to performing the same actions to any complex network of information without being able to recognize the dynamics of its structure and without having adequate support from methods and techniques applied for this purpose. In summary, the characteristics that connect the curriculum data's nature to Big Data theory within the context of undergraduate medical education and as previously analyzed exceed the human abilities to easily process them, are:

- The complexity of both conceptual and actual structure of the curriculum
- The large size of documents and worksheets consisting the curriculum
- The fact that curriculum is accessible from different sources and in different forms and
- The heterogeneity of the curriculum data

This constitutes the main factor that implies the need to find novel ways to reduce complexity of the medical curriculum, transform it to an understandable network of information and render it to a flexible supporting tool in hands of stakeholders. In this way they could be supported to perform analysis of the curriculum and make decisions concerning current and future state of medical education within a given course, easily perceive how learning outcomes are addressed between different courses or for the existence or not of the constructive alignment of the whole program. Additionally, they could be supported to apply in present changes in an effort to alter and improve healthcare education in the future in order to constantly follow the changeable pace of healthcare.

Curriculum Mapping

As we found from literature review a popular way to make sense of a curriculum is to analyze it using the curriculum mapping theory. Harden defines Curriculum Mapping as ‘’*about representing spatially the different components of the curriculum so that the whole picture and the relationships and connections between the parts of the map are easily seen*’’ (Harden

& Association for Medical Education in Europe, 2001). What curriculum mapping indicates as the appropriate way to analyze a curriculum is when looking from the point of directors, teachers, developers and students to be able to answer to questions like what the curriculum covers and how it is assessed, how assessment connects to teaching methods and how students achieve to learn what it is described in the intended learning outcomes. Substantially, this means to be able to distinguish within the curriculum the different entities as described in the introduction section which are what it is taught (content of learning activities), how it is taught (teaching methods), how it is assessed (assessment) and the connections between all of them in order to achieve the intended learning outcomes. Curriculum mapping gives a strong general background on what it is important in a curriculum and using it as guide can be supportive for that purpose. Due to the variety of each educational system and setting and consequently the variety of each individual educational program, curriculum mapping cannot be a ‘’panacea’’ that purely can be applied everywhere. This implies that theory from curriculum mapping must be adapted to the setting of the study and data under investigation.

Visual Analytics

Methods and techniques that can manipulate data in many different disciplines have been developed (Witten & Frank, 2005; Steele & Iliinsky, 2010). Visual Analytics (VA) shown below

in Fig. 1 is a relatively new research field that combines two frequently used

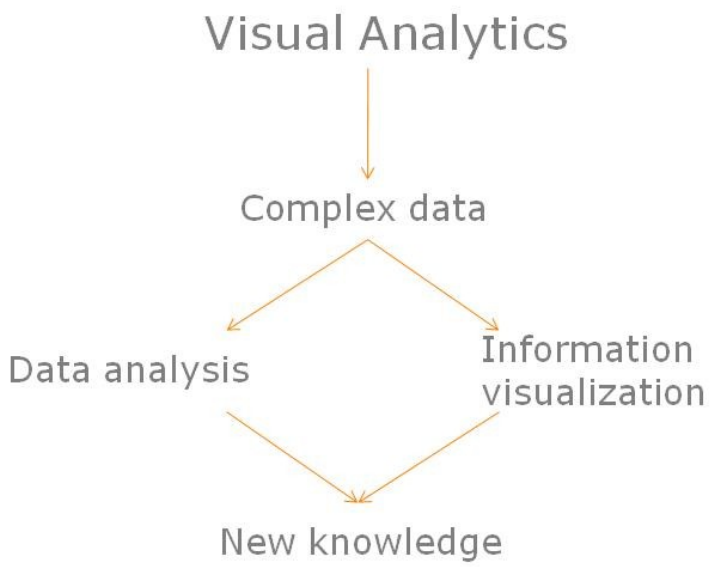

Figure 1. Visual Analytics impact on complex data.

techniques, information visualization and data analysis, along with the ability of human perception (Keim, Mansmann, & Thomas, 2010). Information visualization is ''*the graphical presentation of abstract data*'' which ''*attempts to reduce the time and the mental effort users need to analyze large datasets*'' (Pantazos, 2012). The type of data analysis performed in VA is defined explicitly by the discipline being studied and the nature of the data under investigation. Data analysis for example can be translated into data mining techniques when analyzing and using weather data of previous years to make predictions about the weather in the future or to adjust products in a store according to customers' previous recorded buying preferences (Witten & Frank, 2005). The main purpose of VA is to support the manipulation and exploitation of complicated big data, and create a holistic view of the data, in order to positively impact on analytical reasoning and decision-making (Keim et al., 2010). VA allows the disclosure of previously unknown hidden information and patterns within the data, the use of cognitive strengths like perception and visual pattern recognition, and finally, the presentation of the processed information using visualization techniques (Keim et al., 2009; Steed et al., 2012). In a

healthcare education context Olmos and Corrin reported how analysis and a simple visualization of data, extracted from a healthcare education system, enabled involved stakeholders to instantly review and preview the effects of implemented and future changes (Olmos & Corrin, 2012).

Visual Analytics in Medical Education

The medical curriculum is the main instrument used by different stakeholders to plan, design and deliver healthcare education in a continuous effort to improve it. Due to its complexity and the three characteristics (volume, variety and velocity) described in previous sections it remains unexploited for healthcare education improvement purposes. Additionally, there is lack of empirical data about possible use and benefits of VA in healthcare education.

The purpose of this study was therefore to explore novel ways of analyzing and representing medical curriculum data using visual analytics; in order to (i) identify different aspects which affect how the education is conducted and (ii) use VA to further analyze and visualize the identified aspects using a pilot course from the undergraduate medical program.

Material & Methods

Exploring novel ways of analyzing and representing medical curriculum data implies the creation of an abstract model to represent the curriculum data in the initial form. Therefore this study followed the model methodology because ''*Modeling is the purposeful abstraction of a real or a planned system with the objective of reducing it to a limited, but representative, set of components and interactions that allow the qualitative and quantitative description of its properties*''. In addition, modeling methodology does not concentrate only on the model itself, but allows the model to be used as instrument to study the research object and is not strictly define the modeling approach rather is flexible allowing the researcher to make decisions concerning the importance of the aspects of the real system that will be modeled. (Elio et al., 2011)

Analysis of Collected Data

To build a scientific basis and determine what is important to visualize within the medical curriculum we firstly performed analysis on the collected curriculum data. These data are available in text format and spread in a large amount of different worksheets. They consist of different learning activities (teaching methods), assessment methods (written and other types of

examination), learning outcomes (LO1-LO16) and main outcomes (knowledge, skills and attitude). They summarize the medical curriculum from the different sources and forms that can be found and describe it separately for each course. They also describe it through an overview moving hierarchically from higher education board to actual medical program, to courses within the program and to different parts of a course in multipart courses. These parts of hierarchy from higher education board, to medical program, to program courses and vice versa will be defined from now as levels in medical education/curriculum. Additionally, we mapped the analysis of the collected data to Medbiquitous standards. Therefore, whenever teaching and assessment methods are described at the same time will be referred to as Events and learning outcomes and main outcomes as Expectations. Whenever all of them are described at the same time will be referred as entities. Finally, we considered recommendations of other studies outside the context of undergraduate healthcare education that use same methods in the level of the analysis of the medical curriculum (Corrin & Olmos, 2010; Hege et al., 2010).

After reviewing and analyzing the collected data concerning the whole medical program, we concluded to the Clinical Medicine – Reproduction and Development (CM – RD) course which was selected as a pilot course of the medical program. We chose this course between equally important courses of the medical education; because firstly CM-RD is a semester-duration course resulting in 22,5 credits for students and secondly because it contains more comprehensive information than any other course within the collected curriculum data concerning teaching methods, assessment and learning outcomes. It is a multipart course and consists of the pediatrics/obstetrics and gynecology parts. This study explores the course as a whole and not through the different parts. Therefore the teaching methods are common for both parts, the questions in written examination examine both parts and the learning outcomes are those defined that the students should know after finishing the course and are listed in Appendix S4.

From the chosen course we selected one assessment method (written examination), all the teaching methods and the learning outcomes used in the course for investigation.

While we have applied the presented VA approach to one course of the medical program, it remains flexible and allows to be expanded and applied to other courses of the program and to the whole curriculum in case of a longer type of similar study.

Aspects Identification

We applied curriculum mapping theory on the selected curriculum data from the CM – RD course. According to that, the different entities (written examination, teaching methods and

learning outcomes in our study) and the connections between them when are distinguished and highlighted, they can diagrammatically represent the curriculum as mentioned previously in curriculum mapping section. This helped us to identify for the teaching part of the course, what teaching methods and how are used to address the different learning outcomes, and for the assessment part of the course, how the questions in written examination are used to assess these learning outcomes and consequently the main outcomes. Also, we identified connections between all entities in our data in order to create a holistic view of the course. Thus we identified each entity's role and how it contributes to the overall structure of the course creating paths from a teaching method to an assessment towards the learning and main outcomes. The identification of these aspects led us to finally discern the existence or non-existence of the constructive alignment of the course and to perform gap analysis. The identified aspects and the relations between them are presented in the results section.

Selection of VA Tool

We performed a literature review to identify appropriate tools to manipulate and visualize the identified aspects of the chosen course. However, there are no scientifically validated VA techniques or reported appropriate tools for the analysis and visualization/representation of curriculum data. Therefore, we explored existing tools and techniques from a plethora of open-source and proprietary software already applied for similar purposes of data visualization. We investigated the tools below:

- Microsoft Excel for creating charts of different types where transfer from data to picture is supported efficiently (http://office.microsoft.com/en-001/excel-help/charts-i-how-to-create-a-chart-RZ001105505.aspx)
- Google charts where all popular data representation approaches from bar charts to tree-maps are available and very friendly to use to depict data (https://developers.google.com/chart/)
- Gephi, which ''*is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs*'' (https://gephi.org/)
- BIRT, which is a web-based application and allows the collection of data from multiple sources in order to create visualizations (http://www.eclipse.org/birt/phoenix/intro/) and
- Cytoscape which is ''*a platform for visualizing complex networks and integrating these with any type of attribute data*'' (http://www.Cytoscape.org/).

At the early years of Cytoscape release it was intended to be used for biological research but lately was expanded to be applicable to other disciplines as well. ‘’*Although Cytoscape was originally designed for biological research, now it is a general platform for complex network analysis and visualization*’’ (http://www.Cytoscape.org/). We found Microsoft Excel and Google Charts to be useful to this study in a certain extent as these two tools give a statistical character to the visualizations and allow a unilateral investigation of the data rather than enabling both quantitative and qualitative approaches. Therefore, we excluded both from being used to visualize the curriculum data. Furthermore, BIRT seemed to be a promising tool since allows the manipulation of data with complexity and even from different sources but the data must be first reformed in order to be homogeneous and appropriate for processing by the tool. Since the intention of this study was to perform non-physical transformation on primary data from the CM-RD course this tool was excluded from being used. Finally Gephi and Cytoscape offer the abilities that were more suitable for this study. The philosophy of these two tools is based on complex network analysis and representation. Due to familiarity of the programming language and environment offered in Cytoscape we selected it to proceed and perform the manipulation and visualization of the curriculum data.

Exploring the Medical Curriculum Data

To build the network which will be visualized, Cytoscape uses edge connections between nodes in a network. To do that simple text editors such as Notepad++ (http://notepad-plus-plus.org/) can be used to build the network row by row. This was an ideal approach for this study because allowed the construction of the network accordingly to the identified aspects. For example, to manipulate the entities and represent the connections between them in the CM-RD course, one line of text was Teaching_Method_1-Edge-Learning_Outcome1 which corresponds to Node-Edge-Node representation shown in Fig. 2.

```
TM_Overall edge TM1
TM_Overall edge TM2
TM1 edge LO1
TM1 edge LO5
TM1 edge LO12
TM2 edge LO2
TM2 edge LO5
TM2 edge LO8
Skills edge LO5
Skills edge LO6
Skills edge LO9
Knowledge edge LO1
Knowledge edge LO2
Knowledge edge LO3
.
.
.
.
.
.
.
```

Figure 2. Text file containing part of the network of information before use it in Cytoscape (TM = Teaching Method, LO = Learning Outcome)

The final text file containing the network was subsequently uploaded to Cytoscape which automatically recognizes the form of the network and gives the ability to choose between different shapes, colors and to freely move and rearrange spatially nodes and edges to represent it. We followed this strategy to manipulate the curriculum data and build the networks for three different visualizations. The first visualization is the teaching methods and learning outcomes (taught and non-taught). The second is the written examination questions, the learning outcomes (assessed and non-assessed) and the main outcomes. Finally the third is the constructive alignment consisting of teaching methods, groups of number of points in written examination, the learning outcomes (taught and non-taught and assessed and non-assessed), the main outcomes and the results of students' answers in written examination of the CM-RD course. We designed the three networks to emphasize the role of each entity and the connections between all different entities in order to allow the demonstration of previously perceived and non-perceived patterns and relationships within the curriculum data of the CM-RD course.

Study Framework

An overview of the study framework is presented in Fig. 3. On the left side it is depicted how curriculum data can be currently found in different sources and forms. In the middle (big circle) it is depicted the VA intervention on curriculum data applied in this study and on the right side it is

depicted how curriculum data it is expected to be in a structured form with connections between entities and different courses of the whole medical program being apparent after the intervention of VA.

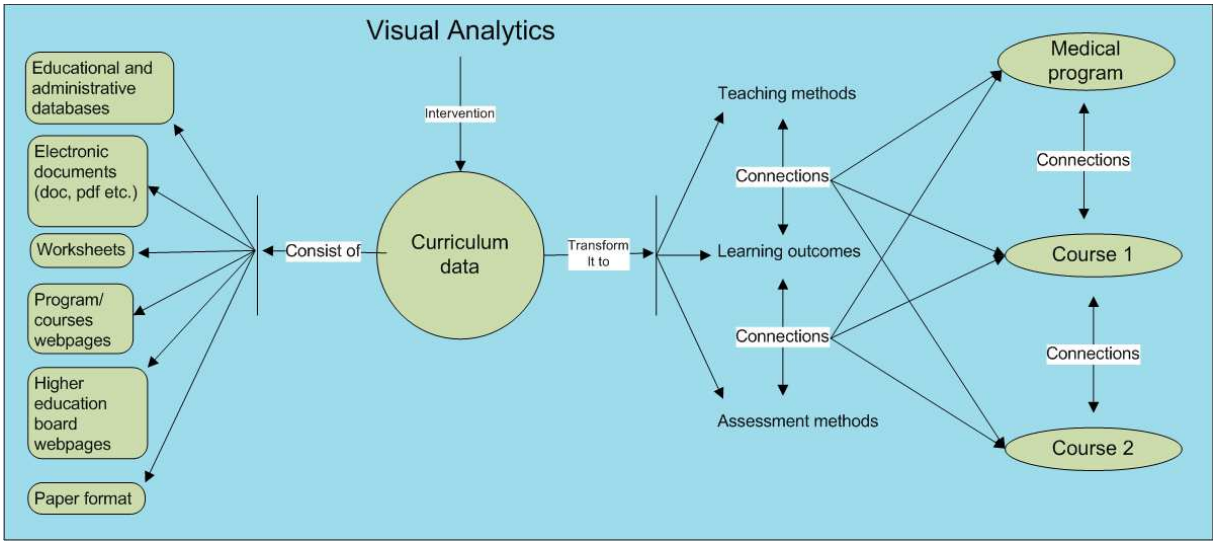

Figure 3. The study framework for analyzing and representing the curriculum data

Results

Identified Aspects

Below are the identified aspects after the analysis of the curriculum data. Learning Outcomes as referred below correspond to the sixteen learning outcomes in Appendix S4 along with Main Outcomes which are Knowledge and Understanding (referred here as Knowledge), Competence and Skills (referred here as Skills) and Judgement and Approach (referred here as Attitude). All those aspects are defined by the Swedish Higher Board of Education as:

1. The teaching methods used in the course (A1)
2. The percentages (proportions) that each teaching method is used (A2)
3. The learning outcomes taught in each teaching method (A3)
4. The percentages that each teaching method uses to address each learning outcome (A4)
5. What learning outcomes are not addressed in any of the teaching methods (A5)
6. The proportion of questions of written examination that assesses each of the learning outcomes (A6)
7. Proportions of learning outcomes correspond to the three main outcomes that are assessed in written examination. (A7)
8. What learning outcomes are not assessed in written examination (A8)
9. Proportions of maximum points of questions in written examination and relation to learning outcomes and main outcomes. (A9)

10 Results of students' answers in written examination and relation to learning outcomes and main outcomes. (A10)
11 The alignment (connections) between teaching methods, written examination, learning outcomes and three main outcomes. (A11)

Learning Outcomes and Teaching Methods

In Fig. S1, the six teaching methods of the course are depicted with green color and the sixteen taught and non-taught learning outcomes are depicted with red color (points A1 – A5 above). The connections between total percentage (100%) of teaching methods and six teaching methods depict the percentages each teaching method is used in the course. The connections between teaching methods and learning outcomes depict the percentages each teaching method's content is used to teach each learning outcome. For example, 2% out of 15% of lectures content is used to teach learning outcome number one (LO:1).

Examination and Learning Outcomes

In Fig. S2, the percentages of the total number of questions (34) used in the written examination are depicted in the connections between blue circle (100% of the questions) and the assessed and non-assessed learning outcomes in red color (points A6 – A8 above). For example, eleven questions (32%) are used to assess LO:5. The learning outcomes are in groups and connected to corresponding main outcomes which are depicted with yellow color. In cases that multiple main outcomes are assessed in group of questions, the total percentage of multiple main outcomes is divided to single main outcomes. For example, 30% of the questions is used to assess skills and knowledge corresponding to 15% skills and 15% knowledge.

Teaching Methods, Learning Outcomes, Examination Results and Gap Analysis

In Fig. S3, teaching methods are depicted with green color, main outcomes with blue color, learning outcomes (taught on the left side of the blue circles, non-taught on the bottom left side, assessed on the right side of the blue circles, non-assessed on the bottom center and assessed but non-taught on the bottom right side) with dark pink color and number of points of questions in written examination are depicted with orange color (points A9 – A11 above). Percentages on connections between assessed learning outcomes and number of points depict the success rate on a learning outcome from an average of sixteen students' answers in written examination. The three dark pink circles surrounded with black line color on the right side of the blue circles,

361 depict the three different places that assessed but non-taught on the bottom right side LO:4 can be
362 found.

363 Discussion

364 This study attempted to use VA to provide novel ways of analyzing and representing big
365 educational data that are regularly collected for healthcare education evaluation purposes. The
366 evaluation of different representations of one chosen course of the medical program produced
367 with VA techniques, show that they have the potential to positively impact on perceiving entities
368 and relations within the curriculum data. For example gaps in learning outcomes which were not
369 previously perceived were identified, revealing shortcomings in the constructive alignment of the
370 examined course. Additionally, the different representations provide with an overview of the
371 course that can be used to plan and apply desired changes in present that could affect healthcare
372 education delivery in the future.

373 Identified Aspects of the Medical Curriculum

374 The analysis of the curriculum data of the undergraduate healthcare education was based on
375 curriculum mapping theory (Harden & Association for Medical Education in Europe, 2001) and
376 on our effort to firstly identify how the CM-RD course is structured through the connections of
377 different entities in different levels as there is no defined concrete series of steps one should take
378 that will lead to a commonly accepted way to analyze and visually represent a medical
379 curriculum (Gathright et al., 2009) and consequently a course of the curriculum. That resulted in
380 identifying aspects as they are listed in results section and building a good understanding of the
381 current structure of the course (Events and Expectations, http://medbiq.org/) and how all entities
382 in it play important role to medical education delivery and quality improvement. Thereby the VA
383 dynamics and possible positive impact on analyzing and representing big educational data were
384 pilot tested and verified in a small scale but still support conclusions for possible usage of this
385 technique in a larger scale to more courses or the whole medical curriculum. Additionally, it
386 establishes a novel way of analyzing and representing a medical curriculum which has the
387 potential to support stakeholders to broadly analyze and make sense of it. An example is the
388 revealed and non-previously perceived discrepancies in the delivery of the medical curriculum
389 (Hege et al., 2010) like the learning outcome (LO4 in Fig. S3) that is assessed in the written
390 examination of the CM-RD course in three different cases but is not taught in any of the teaching

methods. This approach could allow the stakeholders of medical education to deliver a curriculum without gaps preserving thus the desirable constructive alignment in the course and consequently in the curriculum.

Analysis of Pilot Course

The potentials offered by VA (Keim et al., 2009; Steed et al., 2012) and presented in the results making it a promising tool to explore how the Big Data that are regularly collected during the evaluation of healthcare education in Sweden could contribute to the continuous improvement of the healthcare education.

More particularly, the enormous amounts of educational data produced in medical education in relation to teaching, learning, assessment and outcomes and the different sources and forms these educational data can be found, make it an area in which Big Data and Analytics can be very useful to use them to make sense of the complex information to be found in large diverse datasets (Ellawayet al., 2014).

Since there was not found any validated and suggested way of analyzing and representing curriculum data, this study cannot be easily related to other studies. Nevertheless, our findings concerning the demonstrated ability and potentials of VA and selected tool to reduce the complexity of curriculum data and make it a an understandable network of information are in line with a study by Olmos and Corrin who reported how analysis and a simple visualization of data, extracted from a medical education system, allowed involved stakeholders to instantly review and preview the effects of implemented and future changes (Olmos & Corrin, 2012). The rearrangement of nodes and edges representing the entities and connections was made in a small extent intuitively and that can lead to endless tries to best represent it especially if the whole curriculum is to be analyzed and represented. Additionally, in case that an entity or group of entities and connections need to be altered to apply for example in the representations desired changes of the curriculum, a numerous of static images must be created in order to be able to create a comparable before - after picture of part of the curriculum on change. Also, the interactivity allowed from produced static pictures is at low levels. Moreover, based on the analysis of the CM-RD course, the resulted representations (Fig. S1, S2 and S3) show the connections between each teaching method, questions in examination and learning outcomes. Even if that is not the deepest level of analyzing the course as the analysis can go deeper to types of lectures, seminars, clinical training etc. used in the teaching methods and for both parts of the course (pediatrics and gynecology), it can be complex enough to represent all these details. If all

these details then need to be merged from different courses to represent the whole curriculum, the produced network can be extremely complex to be perceived.

On the other hand the rearrangement of nodes and edges resulted in different views of the same inserted network giving thus the ability to add one very important layer in the analysis of the curriculum data. This additional layer promotes the intuitiveness of the researcher to produce the different representations having as base established theory of curriculum data analysis from curriculum mapping theory. In this manner, the resulted representations (Fig. S1, S2 and S3) indicate that this approach of analyzing and visualizing the course brought positive results and opens a new way of viewing at the CM-RD course and potentially to complex medical curriculum. Even if a snapshot of the whole curriculum was used to produce the representations they indicate additionally that they have the potential to achieve in transforming the complex information of the medical curriculum into a structured and comprehensible network.

Representation of Pilot Course

The selected tool for analyzing and representing the CM-RD course was Cytoscape. Even though that between the different explored tools in this study Cytoscape was the one that suited better for the analysis and representation of the curriculum data, it can produce only static images of networks allowing thus low levels of interaction with the resulted representations.

It also requires a lot of effort and familiarity with similar software to build the networks (Fig. 2) before they can be entered and recognized from Cytoscape. This applies even for only one course of the curriculum as in this study.

On the contrary, a big advantage of Cytoscape is that allows the user to create easily multiple representations of the same network and support intuition and high levels of analysis before the choice of final representations. The potentials offered by Cytoscape were considered as appropriate for this study's purpose as it brought into light facts that were not perceived until now like the gaps in the structure of the course or the disproportionate way that skills, knowledge and attitude are assessed in the written examination. Therefore, it establishes Cytoscape as an appropriate VA tool to represent the identified aspects of the CM-RD course in comparison to all other tested tools.

Strengths and Limitations

The main strength of the study was the access to genuine educational data used currently in medical education. These data were prepared for review from the Swedish Higher Education

Authority, summarize the medical curriculum and after reviewed, considered as appropriate to conduct this study. Also the model methodology followed in this study added flexibility in terms of deciding which aspects of the medical curriculum should be modeled and visualized to create the final model.

On the other hand this study was limited in analyzing one course's structure and more particular the teaching methods, the written examination and the relation between them and to the learning and main outcomes.

Implications for Healthcare Education

The findings of this study contribute to the medical education field with new knowledge using VA. They also open a new area for investigation in medical education informatics field. The VA approach used in this study to analyze and represent the CM – RD course through the different representations seems to have good potential to:

- Reveal the hidden structure of the examined curriculum data between different entities
- Identify gaps and roles of minor-major entities in the structure of the data
- Possibly used as instrument for planning and apply future changes in the curriculum in present in an effort to be able to constantly align medical education with demands of changeable healthcare setting.

More general the contribution of this study resides on the novel ways that provides on verifying ongoing medical education structure and analyzing and deciding about design and plan of activities in hands of teachers and directors to support medical education improvement. Since Swedish Higher Education Authority defines the learning outcomes for undergraduate medical education at national level it would be interesting to apply the same VA techniques presented in this study to medical education in other places in the country and investigate possible similarities and differences between them towards the study objectives. Nevertheless, to reach this point, that medical curricula from different regional universities can be compared even in the level of a pilot course, significant preparation work must be undertaken. As explained in introduction section, the data used in this study summarize the data from medical curriculum as they exist in different places and forms. This means that the preparation work from unstructured to structured - but still complex and unexploited for medical education improvement purposes - curriculum data, had already been done. These preparation steps require devoting time, resources and effort and apply expertise to produce data that can be further analyzed and presented with VA. This adds another

"thick" layer of processing raw big educational data but as this study demonstrates, the added value of using VA constitutes a great justification to fire up such initiatives.

The fact that VA techniques must be applied in the data used in this study implies that, even if they are not in the unstructured form as their primary curriculum data they remain big educational data but in another category. As analyzed in Big Data section the three characteristics that coexist in data possessed in a system or domain are enough to challenge constrains to manipulate and analyze them so they can be used. Then within this system or domain these data are considered as Big Data irrespective of whether they can be considered 'small' to another domain. Depending on the domain the volume of data can vary from megabytes to petabytes. For example a 40MB presentation is considered big in comparison to typical size of a PowerPoint presentation. Thus, Big Data may refer to different sizes and types from domain to domain but all these domains share a common challenge that must cope with, and is to being able to search, analyze and make sense of the data. (Zaslavsky, Perera, & Georgakopoulos, 2013)

The separation in different categories of big educational data is derived from the fact that techniques that applied in the data of this study cannot be applied in the primary data and vice versa. Without the use of special techniques to analyze and represent the summarized data they would remain unexploited as they are until today for medical education improvement purposes because of their complexity. Different techniques applied for different purposes and in the case of primary data they must be formed so they can be further processed with VA with different techniques than the one demonstrated in this study and this is not described as it is outside this study's boundaries. Additionally, this study has the potential to be generalized outside Swedish boundaries. The current data were analyzed partially based on existing curriculum mapping theory that has been broadly used in higher education in general and more extensively in medical education.

Future Research

The findings of this study suggest that further investigation is required towards both directions for reducing the complexity of the whole medical curriculum and deeper to the different parts of multipart courses to create a holistic view of medical education and be able to draw conclusions that can affect it from a more general view. To achieve this, new, more interactive ways of representing the curriculum with more details and at the same time reducing its complexity, must be investigated. This involves investigating new tools that are able to perform such actions or the creation of customized tools for these purposes.

Also, the current approach must be adjusted to analyze and represent multipart courses in healthcare education with more details without increasing the complexity of representations to unacceptable levels.

Conclusions

In this study we explored the potential of Visual Analytics to identify, analyze and represent big educational data that are regularly collected for healthcare education evaluation purposes. Through eleven different aspects which affect how the medical education is conducted and an abstract model of the examined data in three different variants we evaluated a course and concluded that Visual Analytics could be used to reveal novel ways of representing medical educational curriculum. This finding could have positive implications on medical education informatics research and on how quality improvement of medical education is designed.

Acknowledgments

We wish to thank all the staff at Karolinska Institutet, Sweden that provided the authors of this study with assistance, comments and encouragements. This study was funded with intramural grants.

References

Biggs, J., & Tang, C. (2007). Teaching for quality learning at university (Society for Research into Higher Education).

Cho, N., Gilchrist, C., Costain, G., & Rosenblum, N. (2011). Incorporating evidence-based medicine in the undergraduate medical curriculum: Early exposure to a journal club may be a viable solution. *UTMJ, 88*(3), 154-155.

Corrin, L., & Olmos, M. (2010). Capturing clinical experiences: supporting medical education through the implementation of an online Clinical Log.

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., & Zikopoulos, P. (2012). Understanding big data: McGraw-Hill.

Elio, R., Hoover, J., Nikolaidis, I., Salavatipour, M., Stewart, L., & Wong, K. (2011). About Computing Science Research Methodology.

Ellaway, R. H., Pusic, M. V., Galbraith, R. M., & Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical teacher, 36*(3), 216-222.

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., . . . Kelley, P. (2011). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Revista Peruana de Medicina Experimental y Salud Pública, 28*(2), 337-341.

Gannod, B. D., Gannod, G. C., & Henderson, M. R. (2005, 2005). *Course, program, and curriculum gaps: assessing curricula for targeted change*.

Gathright, M. M., Thrush, C., Jarvis, R., Hicks, M. E., Cargile, C., Clardy, J., & O'Sullivan, P. (2009). Identifying areas for curricular program improvement based on perceptions of skills, competencies, and performance. *Academic Psychiatry, 33*(1), 37-42.

Harden, R. M., & Association for Medical Education in, E. (2001). *Curriculum mapping: a tool for transparent and authentic teaching and learning*: Association for Medical Education in Europe.

Hege, I., Nowak, D., Kolb, S., Fischer, M. R., & Radon, K. (2010). Developing and analysing a curriculum map in Occupational-and Environmental Medicine. *BMC medical education, 10*(1), 60.

Keim, D. A., Mansmann, F., Stoffel, A., & Ziegler, H. (2009). Visual analytics *Encyclopedia of Database Systems* (pp. 3341-3346): Springer.

Keim, D. A., Mansmann, F., & Thomas, J. (2010). Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter, 11*(2), 5-8.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Maojo, V., Martin, F., Crespo, J., & Billhardt, H. (2002). Theory, abstraction and design in medical informatics. *Methods of information in medicine, 41*(1), 44-50.

Mennin, S. (2010). Self-organisation, integration and curriculum in the complex world of medical education. *Medical education, 44*(1), 20-30.

Olmos, M., & Corrin, L. (2012). Academic analytics in a medical curriculum: Enabling educational excellence. *Australasian Journal of Educational Technology, 28*(1), 1-15.

Pantazos, K. (2012). Custom Visualization without Real Programming.

Picciano, A. G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. *Journal of Asynchronous Learning Networks, 16*(3), 9-20.

Ritko, A. L., & Odlum, M. (2013, 2013). *Gap Analysis of Biomedical Informatics Graduate Education Competencies*.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications, 33*(1), 135-146.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30-32.

Steed, C. A., Potok, T. E., Patton, R. M., Goodall, J. R., Maness, C., & Senter, J. (2012, 2012). *Interactive Visual Analysis of High Throughput Text Streams*.

Steele, J., & Iliinsky, N. (2010). *Beautiful Visualization: Looking at Data through the Eyes of Experts*: " O'Reilly Media, Inc.".

West, D. M. (2012). Big data for education: Data mining, data analytics, and web dashboards. *Governance Studies at Brookings, September.* *http://www.brookings.edu/~/media/research/files/papers/2012/9/04%20education%20technology%20west/04%20education%20technology%20west. pdf*.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. *arXiv preprint arXiv:1301.0159*.