Visualization of the IMIA Yearbook of Medical Informatics Publications over the Last 25 Years

D. W. Yergens[1], H. Tam-Tham[1], E. P. Minty[2]

[1] Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada

[2] Department of Medicine, University of Calgary, Calgary, Alberta, Canada

Summary

Background: The last 25 years have been a period of innovation in the area of medical informatics. The International Medical Informatics Association (IMIA) has published, every year for the last quarter century, the Yearbook of Medical Informatics, collating selected papers from various journals in an attempt to provide a summary of the academic medical informatics literature. The objective of this paper is to visualize the evolution of the medical informatics field over the last 25 years according to the frequency of word occurrences in the papers published in the IMIA Yearbook of Medical Informatics.

Methods: A literature review was conducted examining the IMIA Yearbook of Medical Informatics between 1992 and 2015. These references were collated into a reference manager application to examine the literature using keyword searches, word clouds, and topic clustering. The data was considered in its entirety, as well as segregated into 3 time periods to examine the evolution of main trends over time. Several methods were used, including word clouds, cluster maps, and custom developed web-based information dashboards.

Results: The literature search resulted in a total of 1210 references published in the Yearbook, of which 213 references were excluded, resulting in 997 references for visualization. Overall, we found that publications were more technical and methods-oriented between 1992 and 1999; more clinically and patient-oriented between 2000 and 2009; and noted the emergence of "big data", decision support, and global health in the past decade between 2010 and 2015. Dashboards were additionally created to show individual reference data, as well as, aggregated information.

Conclusion: Medical informatics is a vast and expanding area with new methods and technologies being researched, implemented, and evaluated. Determining visualization approaches that enhance our understanding of literature is an active area of research, and like medical informatics, is constantly evolving as new software and algorithms are developed. This paper examined several approaches for visualizing the medical informatics literature to show historical trends, associations, and aggregated summarized information to illustrate the state and changes in the IMIA Yearbook publications over the last quarter century.

Keywords

Medical Informatics, Visualization, Bibliometrics

Yearb Med Inform 2016;Suppl1:S130-8
http://dx.doi.org/10.15265/IYS-2016-s003
Published online June 30, 2016

Background

The last 25 years have been a period of innovation in the area of medical informatics. Medical informatics is a large and diverse field that involves many other domains focusing on biomedical, clinical, nursing, public health, and other multidisciplinary groups. Even within these groups, the literature is wide ranging, consisting of descriptions of research, implementation, education, evaluation, and assessment of various technologies and methodologies.

For the last quarter century the International Medical Informatics Association (IMIA) has published the Yearbook of Medical Informatics, collating selected papers from various journals in an effort to provide an aggregated summary of the academic literature over the previous year. Every year, the series contains a special theme, as well as a set of general themes that remain relatively consistent from year to year. For example, in 2001 the special section was "Digital Libraries and the Web" [1] and the six other general themes that were included were: Health and Clinical Management; Computer-based Patient Records; Information Systems; Image and Signal Processing; Knowledge Processing and Decision Support Systems; and Education. In 2015, the special theme was "Patient-centered care coordination" [2] and the general themes included were: Health and Clinical Management; Human Factors and Organizational Issues; Health Information Systems; Sensor, Signal and Imaging Informatics; Decision Support; Knowledge Representation and Management; Education and Consumer Health Informatics; Bioinformatics and Translational Informatics; Clinical Research Informatics; Natural Language Processing; and Public Health and Epidemiology Informatics. The last two sections were recently added to address increased interest and specialization in these areas [3].

The longevity and consistency of the IMIA Yearbook of Medical Informatics has allowed the research in this area to be documented and collated, providing the opportunity to explore trends in the literature. Various approaches can be utilized for exploring and summarizing the contents of the given literature around a particular topic. For example, DeShazo [4] reported on publication trends in the medical informatics literature over a 20 year time period by utilizing MeSH terms. Also, Mihalas [5] reported on the evolution of trends in the European Medical Informatics literature spanning over 50 years. Five major periods were identified and labeled: 1950-1975 "early" medical informatics (pre-organization/pioneering); 1975-1990 "childhood/youth" of medical informatics; 1990-2000 "consolidation period"; 2000-2010 "maturity of medical informatics"; and 2010-2020 "full integration" of medical informatics in medicine and healthcare. In terms of visualization, several approaches have been applied towards the graphical display of the literature. Synnestvedt explored several approaches in the mid-2000s, including an investigation of trends within the medical informatics literature [6] and an examination of visualization and knowledge discovery in bibliographical databases through CiteSpace II [7]. Another approach was PubNet[8], which was a web-based system that returned PubMed queries and mapped them into graphical networks.

The objective of this paper is to visualize the evolution of the medical informatics field over the 25 years based on the analysis of the frequency of word occurrences in the papers published in the IMIA Yearbook of Medical Informatics. To accomplish this task, we employed reference management

software for summarizing and aggregating themes from the literature, and explored several visualization techniques including word clouds, topic clustering, and the development of literature-specific dashboards using a variety of charting methods.

Methods

Literature Search

We searched PubMed using the search term: ""Yearbook of medical informatics" [Journal])" to retrieve the IMIA Yearbook of Medical Informatics references from 2006 to 2015. From 1992 to 2005, the table of contents from hard copies of the IMIA Yearbook of Medical Informatics books were used and manually entered into the BibTeX format. The table of contents provided information on the year, authors, title, page numbers, and original article reference. We then searched PubMed for the abstract from the original article reference and included this additional information into the BibTeX reference when possible.

All of the references were then imported into a custom-written Java-based literature review application, hereafter referred to as Synthesis[9], for improved bibliographical management and information extraction. The Synthesis software allows for the advanced searching of words and phrases in the title and abstracts and for associated tagging/annotating of the references based upon keywords, and has been described in detail elsewhere [10].

Theme Identification

We identified major themes in the assembled literature using several approaches and technologies. The first of these approaches was to utilize word clouds by identifying words that appeared frequently in the references (Figure 1). Word clouds are a visualization technique for displaying frequently used words within a body of literature based upon textual data [11] (the technique excludes common words such as articles, prepositions, and conjunctions). The size (font) of the words is adjusted based upon their frequency within the studied set of documents to represent word usage. We implemented a custom-written Java word cloud application in the Synthesis software with a slider bar that increased or decreased the number of words (based upon their frequency) presented in the word cloud. As themes emerged, these were tagged with the associated references.

The second approach we used to explore themes was by grouping similar documents together through automated document clustering. References from Synthesis were accessed through the underlying Lucene database and document clustering was applied using the Carrot2 software (version 3.9.2) [12] with the Lingo algorithm [13] with default settings for theme identification.

The last approach to theme identification was to tag literature references within Synthesis, based on the themes that emerged from previous word clouds and document clustering approaches. This was accomplished through the use of keyword searching within Synthesis from the titles and abstracts using keywords and phrases, Boolean operations (e.g. AND, OR and NOT) and proximity searching. Due to the nature of this literature search, where references may have had multiple themes, several columns were created to capture this information. For example, a reference may have been on the evaluation of neural networks, hence the themes of evaluation and neural networks would be identified as two complimentary concepts within the reference.

Visualization

After identifying themes, we used several approaches to visually display the information. As explained previously, we utilized the word clouds feature in the Synthesis application. Two different visualizations were used to

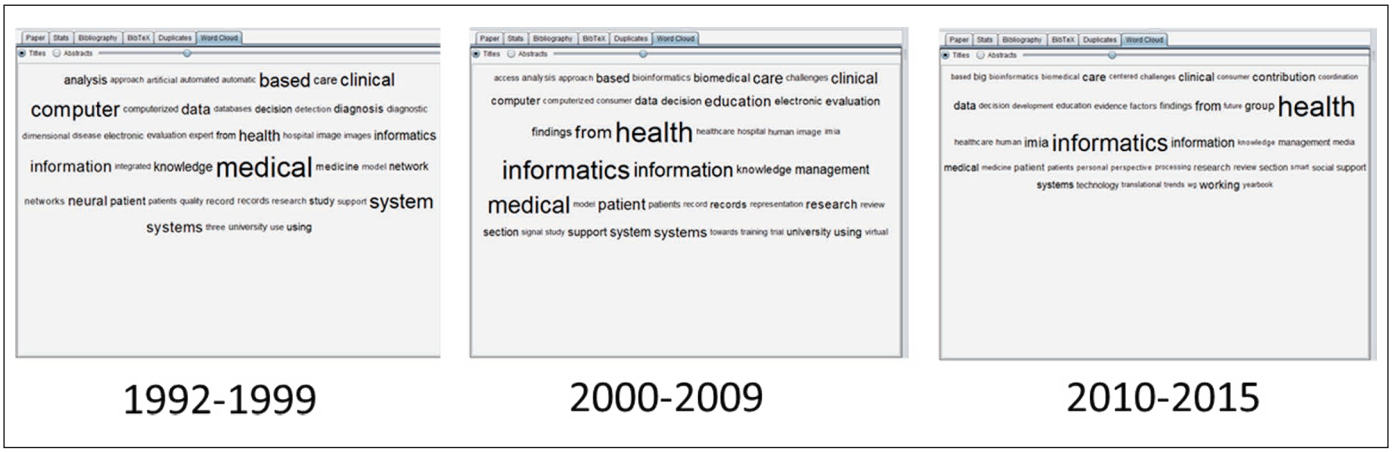

Fig. 1 Word Clouds based on Titles from the IMIA Yearbook of Medical Informatics Publications

view the automatically categorized topics within the Carrot2 application: Foam trees (not shown) and Aduna cluster maps. Foam trees are similar to tree maps and are used to illustrate groupings in the data, but in a non-rectangular format.

The Aduna cluster map displays groupings of data similar to a network graph, but has the advantage of showing how these groups overlap with each other (Figures 2a-2c). Each yellow node represents an IMIA publication with the edges (presented in various colors) representing a theme. Nodes not linked to the network are considered isolates (e.g. "Other topics" and "Computerized Three-dimension" in Figure 2a). Nodes linked to other nodes in the network and located on the periphery of the network represent publications with at least one term that is found in another publication, though relatively fewer overlapping terms than nodes located at the center of the network (e.g. "National Library of Medicine" and "Neural Network"). Nodes located closer to the center of the network have relatively more overlapping terms with other publications (e.g. "Patient Outcomes" and "Clinical Trials").

The next step was the development of the dashboards. A custom-written Java-based application [14] was developed that used the references and associated derived data from Synthesis to create web-based dashboards for displaying the literature information. Several visualization techniques were considered and explored during the development of the dashboard (Figure 3). These included a timeline approach for displaying individual references (title and abstract) based upon publication date (top display of Figure 3); the dashboard displaying (from the bottom left of Figure 3) a tree map for presenting themes and their relation within publication year; a bubble chart for displaying themes based upon year; a flare chart for displaying linkages between authors; a co-occurrence matrix for displaying associations between themes; a choreograph for displaying country information (not shown); and standard charting techniques (e.g. bar, line, stacked area, etc.) for showing aggregated themes across years. The java-script visualization frameworks D3.js [15] and MIT SMILE timeline widget [16] were used in the development of the dashboards.

The dashboards were developed and displayed using two types of technologies. The first technology used was a 39 inch 3840x2160 pixel resolution monitor (4K display). The second technology used was a Visualization Wall located at the University of Calgary Taylor Family Digital Library [17]. This visualization wall measures 16x6 feet, providing a resolution of 9600x3600 pixels.

Results

The literature search resulted in a total of 1210 references, 318 of the references were from PubMed and 892 references were manually entered references from BibTeX. We excluded 213 references, as they were not original research articles (e.g. prefaces, information on IMIA, and various introductory overviews of theme headings). This resulted in 997 references, which we included in our exploration of the published literature.

As explained in the method sections, three visualization approaches were used to identify the major themes and trends in the literature. The first of these was the application of word clouds using the Synthesis application to dynamically look at the frequency of titles and abstracts (Figure 1). The second was the use of the Carrot2 document clustering application, specifically the Aduna cluster map visualization technique. Aduna cluster maps and the associated time intervals can be seen in Figures 2a to 2c. The final approach explored was the presentation of dashboards (Figure 3).

We examined the literature specifically through three different time periods: 1992 to 1999 (404 references), 2000 to 2009 (399 references), and 2010 to 2015 (194 references).

Period: 1992 to 1999

The first period (1992-1999) of the analysis, as illustrated in Figure 2a, had major themes emerge around "Information System" (38 nodes), "Medical Records" (22), and "Hospital Information System" (10) with "Decision Support" (15). Note: the numbers in parentheses represent the number of references associated with each theme from the Aduna cluster maps depicted in Figures 2a-c. The use of more advanced analytical methods such as "Neural Network" (30) and "Artificial Neural Networks" (12) figured prominently. This was complimented with technical evaluation methods such as "Diagnostic Accuracy" (7), "Test Set" (7), and "Predictive Value" (2).

In addition to the technical evaluation methods, there appeared to be themes emerging to assess the "Clinical Use" (23) such as through the use of "Clinical Trials" (8), "Comparative Study" (17), and "Prospective Evaluation" (3). It was also observed that references from 1992-1999 appeared more heterogeneous than the other periods, as themes in this period did not considerably overlap with the other themes.

Period: 2000 to 2009

This second period of the analysis (2000-2009), as illustrated in Figure 2b, showed the importance of "Clinical Data" (38), which was further emphasized with the additional themes of "Patient Records" (31) and the "Electronic Patient Record" (17).

This period also illustrated the development of more system-oriented themes, with an emphasis on systems integration and standardization as seen with the following themes: "System Design" (37), "Development and Evaluation" (29), "Systems Integration" (32), and "Data Fields" (19). This was further emphasized with the theme "Decision Support" (26) and a focus on their efficacy as seen through the themes: "Impact on Clinical" (11), "High Quality" (5), and "Errors" (10). During this period, an appreciable number of references were associated with "Public Health" (18).

Period: 2010 to 2015

The final period (2010-2015) of the analysis as illustrated in Figure 2c showed the emergence of "Big Data" and "Personalized Medicine" as major categories. They were represented through the following associated themes: "Big Data" (19), "Big

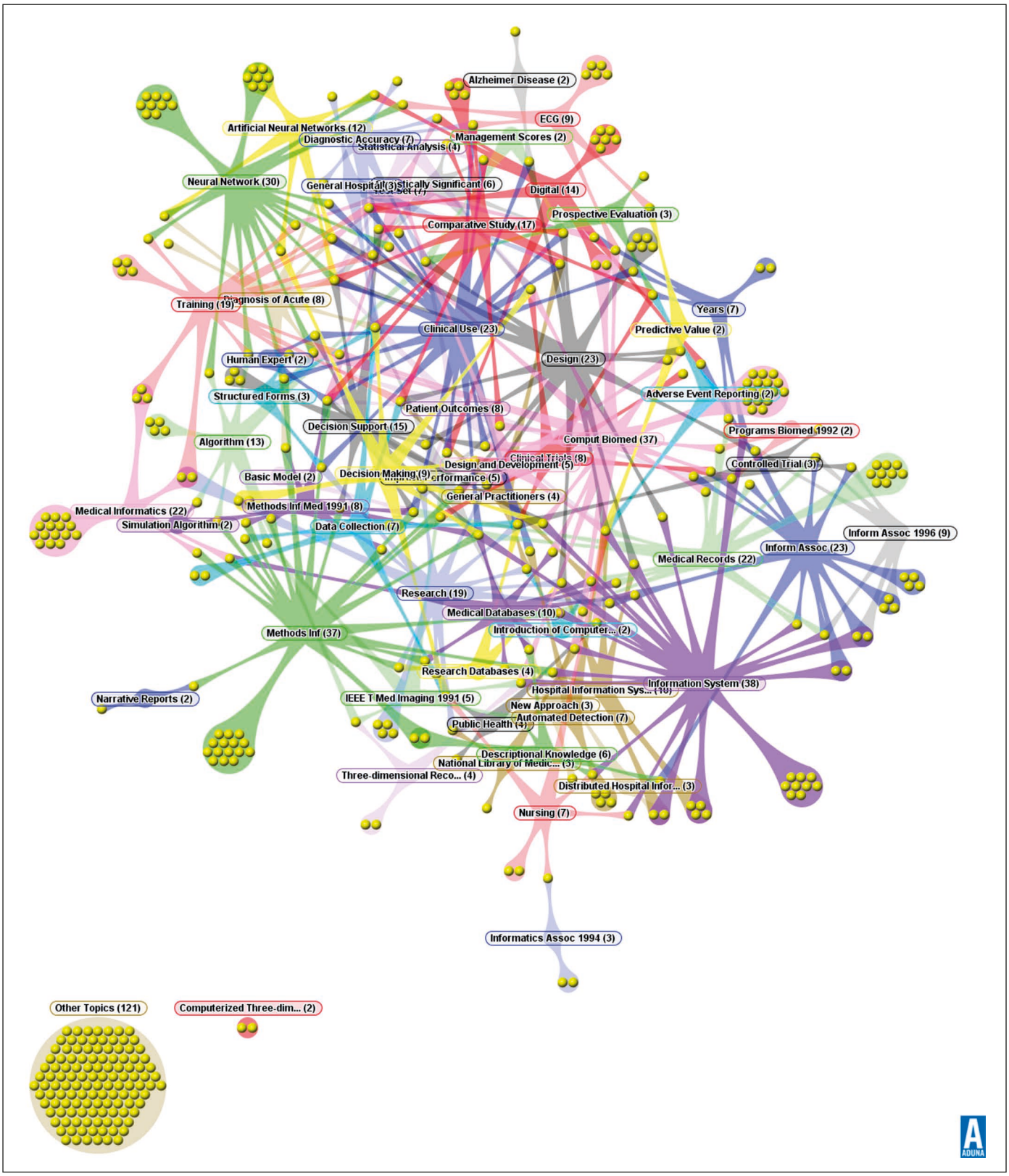

Fig. 2a Aduna Cluster Map of IMIA Yearbook of Medical Informatics Publications in the 1990s (1992 to 1999). Each yellow circular point represents a single publication, grouped by non-mutually exclusive themes determined by the Lingo algorithm in Carrot2 software.

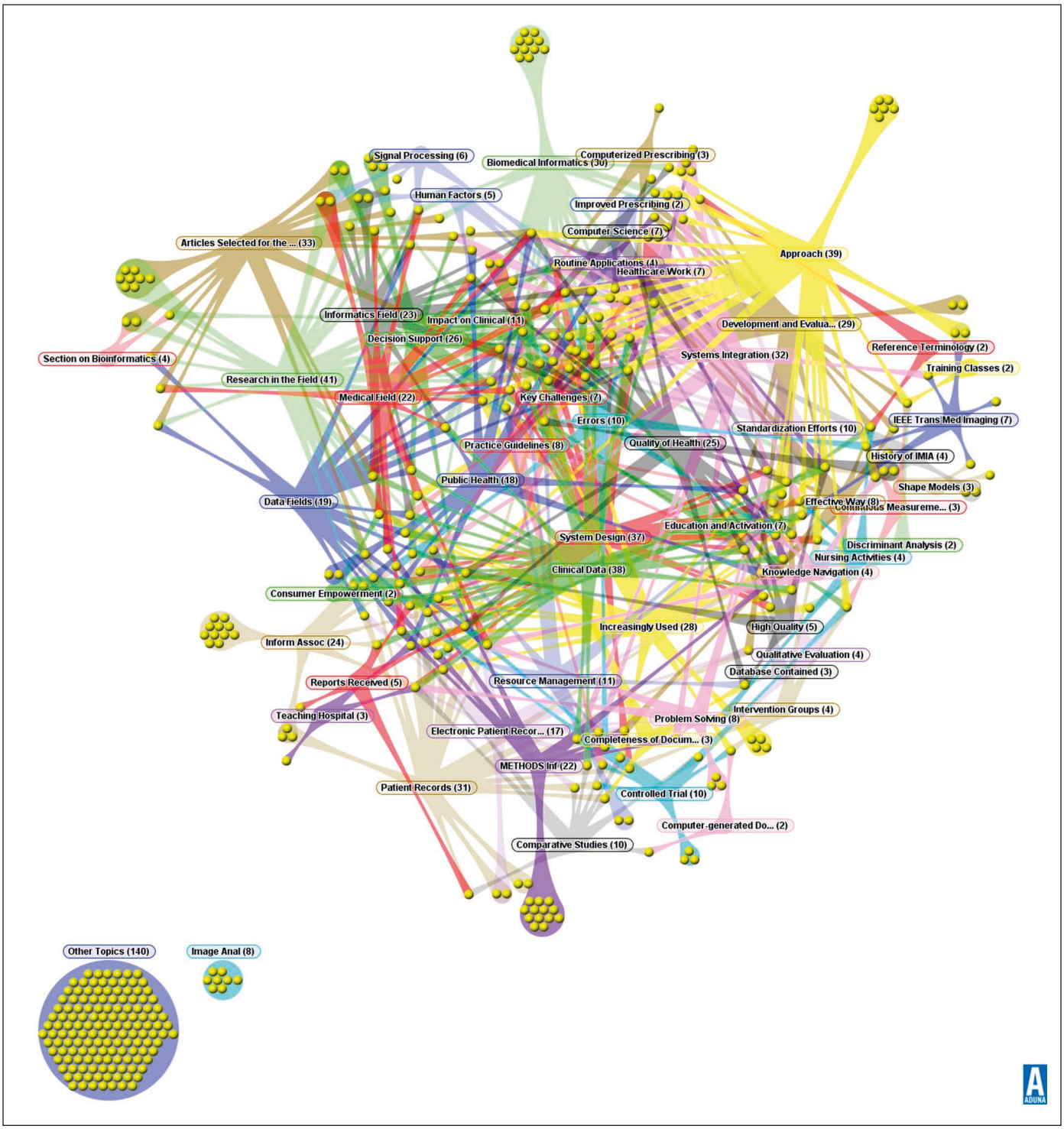

Fig. 2b Aduna Cluster Map of IMIA Yearbook of Medical Informatics Publications in the 2000s (2000 to 2009).

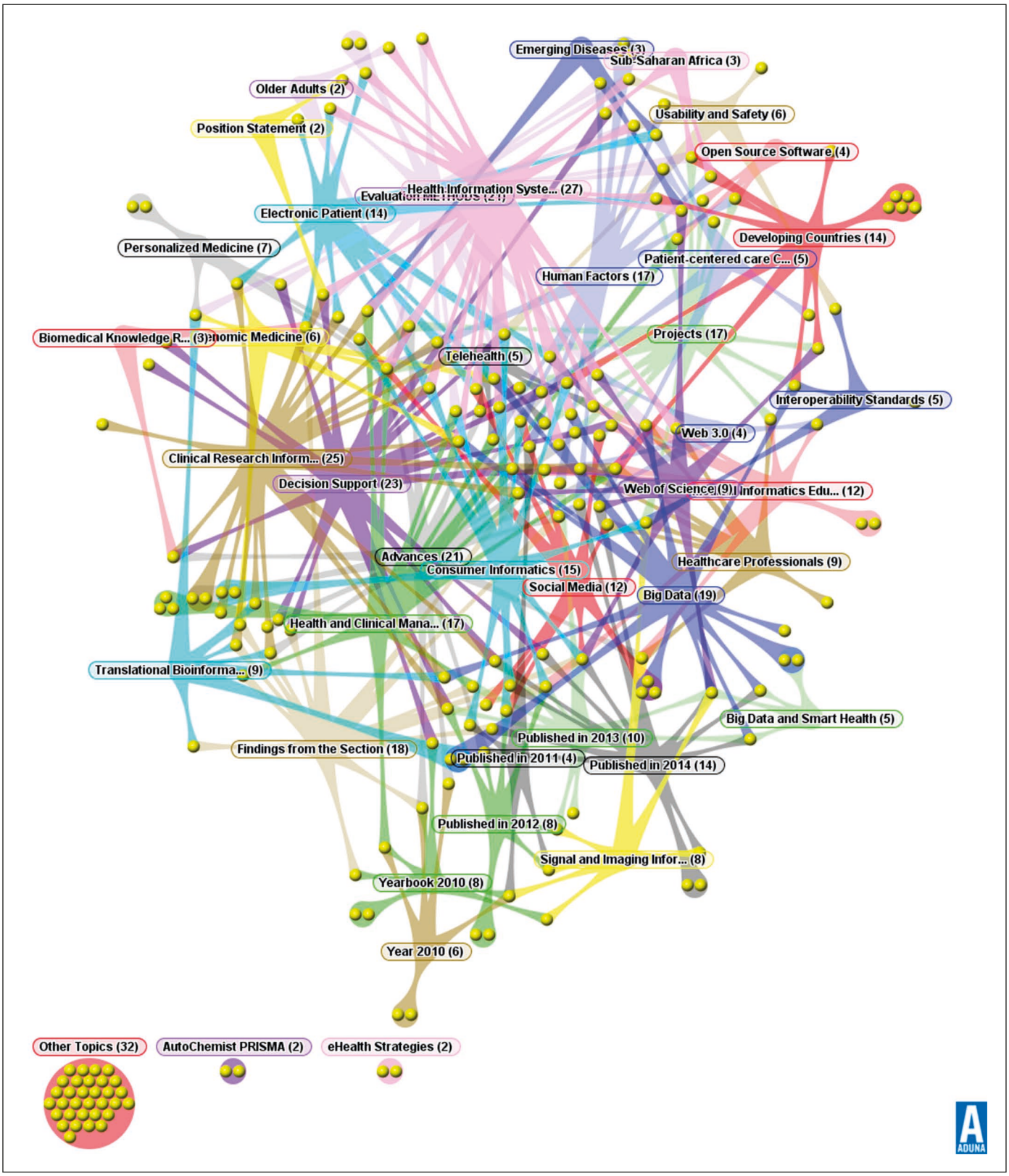

Fig. 2c Aduna Cluster Map of IMIA Yearbook of Medical Informatics Publications in the last five years (2005 to 2015).

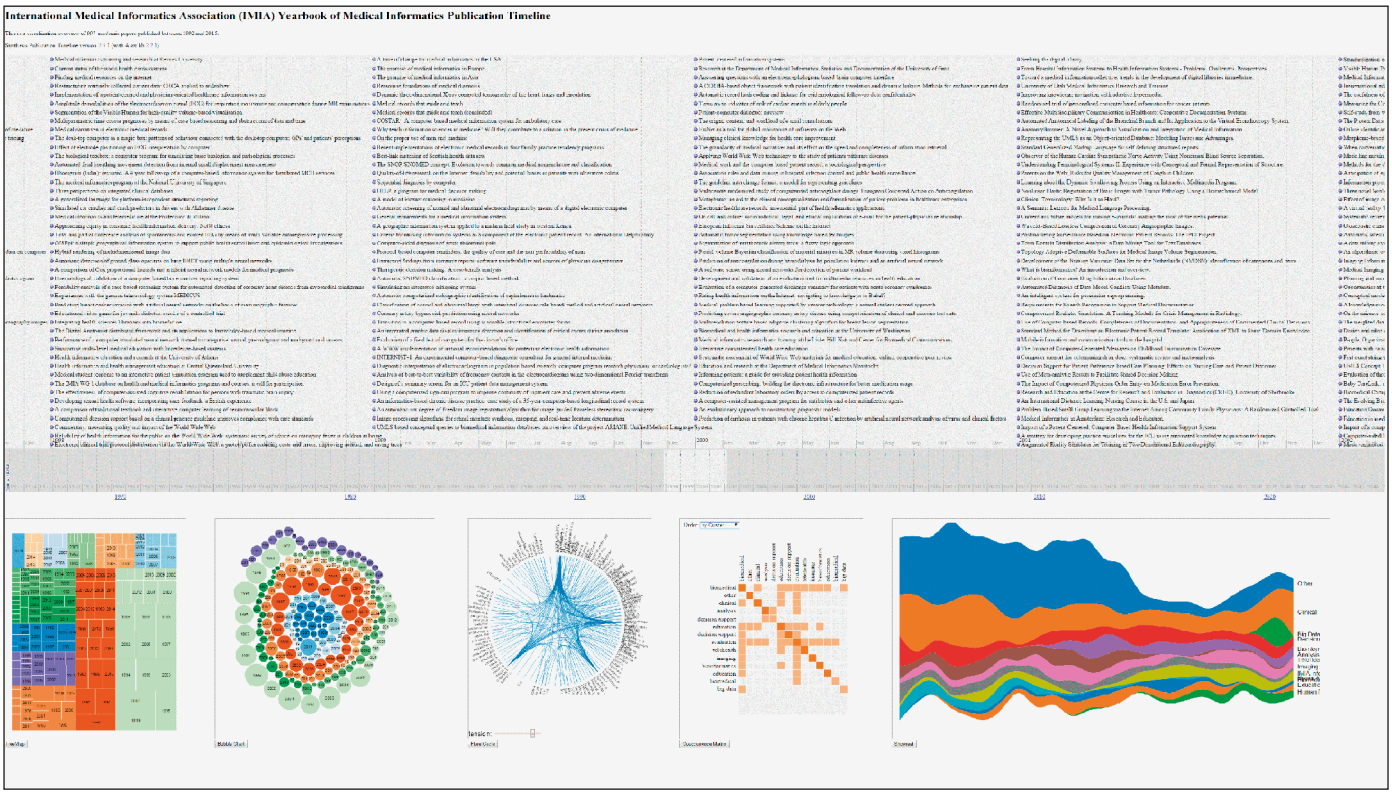

Fig. 3 Example of web-based Dashboard displaying IMIA Yearbook of Medical Informatics publication information.
Top: timeline for displaying individual references (title and abstract) based upon publication date. Bottom from left to right: tree map displaying themes within publication year; bubble chart displaying themes with publication year; flare chart displaying associations between authors; co-occurrence matrix displaying associations between themes; dynamic charting (e.g. bar, line, stacked area, etc.) for showing aggregated themes across publication year.

Data and Smart Health" (5), "Personalized Medicine" (7), "Genomic Medicine" (6), and "Translational Bioinformatics" (9). However, it should be noted that the majority of the references associated with "Big Data" occurred in 2014.

As with the other periods, the themes of "Decision Support" (23), "Health Information Systems" (27), and "Health and Clinical Management" (17) were well represented. This period also saw a rise in the end-user interface of medical informatics through the themes "Human Factors" (17) and "Usability and Safety" (6). Adding to this, themes for supporting stakeholders emerged through the use of "Consumer Informatics" (15), "Social Media" (12), and "Web 3.0" (4). Another category represented was around global health with the following themes emerging: "Developing Countries" (14) and "Sub-Saharan Africa" (3).

Additionally, there appeared to be more overlapping themes among references in the last five years compared to the first decade of literature, suggesting an increase in the number of projects involving multiple methods/techniques in order to address increasingly complex research objectives.

Discussion

In this paper, we explored various approaches for visualizing the last 25 years of the biomedical informatics field using the IMIA Yearbook of Medical Informatics. We utilized word clouds, topic clustering, and information dashboards consisting of several types of charting. We found that the categorization of the literature via word clouds and topic clustering visualization tools were important methodological tools at this early stage to describe the data, while the subsequent dashboards are informative in other aspects. There is no definitive approach for visualizing this kind of information and differing approaches should be utilized to provide a more comprehensive examination of the literature.

As DeShazo [4] indicated, medical informatics is multidisciplinary and heterogeneous, and due to these factors, it is difficult to categorize or collate this area of research. However, the use of cluster mapping has offered an opportunity to inductively generate main trends of the evolution of medical informatics over time.

Our analysis and interpretation of the various visualization approaches appeared to show that the publications were more technical and methods-oriented between 1992 and 1999. This is perhaps a reflection

of an emerging field of research, which was unknown previously, and therefore, open to greater experimentation.

The literature then appeared to become clinically and patient-oriented between 2000 and 2009 with more evaluation-focused publications taking place. Finally, between 2010 and 2015, we see the emergence of the concepts of big data and personalized medicine emphasizing the use of the vast amounts of data that have been generated through the deployment of clinical information systems over the years, as well as the advent of high throughput translational measurement. Additionally, we observed more of a display of global health literature perhaps indicating that the research has reached a level maturity where the approaches can be applied to other environments.

Our interpretation of the evolution of medical informatics parallels many of Mihalas [5] conclusions. Between 1990 and 2000, a consolidation period occurred where medical informatics became less of an "independent discipline". The medical informatics field subsequently matured between 2000 and 2010, with a clearer focus on understanding the potential benefits of e-Health. And finally, in the 2010 to 2015 period, similar to Mihalas' findings, there was a new attention to "big data", "patient empowerment", and the inclusion of genetic data into the electronic health record that was evident from the literature.

In terms of the technical approaches used for this analysis, we found that the generation of meta-data, such as themes, was an essential component in understanding the literature. Word clouds and topic clustering visualizations were important tools at this early stage for helping in grouping and aiding in the initial understanding of the literature. It should be noted that even though the papers were categorized in the IMIA Yearbook of Medical Informatics in themed sections, this information is lost through the bibliometric archiving process, where only some information specific to the paper (i.e. title, authors, and abstract) is recorded and not the sectional theme that the paper is associated with.

However, the theme of the paper is only one aspect that may be important, and the interpretation of the themes may vary between audiences. A reviewer of the literature may be interested in a specific sub-set of the literature such as evaluations, while others may be more interested in the initial introduction of technologies or methods in the body of literature. Therefore, the ability to use other methods of summarizing the literature based upon specific user's interests and objectives is essential.

Future work would benefit from a consensus approach to our field which could be reflected through ontologies that describe informatics and methodological domains. We concede that word frequencies may underestimate occurrence of parent concepts when these give rise to broad sets of related child concepts in the literature. This would also allow visualization at different levels of the hierarchy, and aid in the exploration of orthogonal fields that co-occur in informatics projects (for example, the use of machine learning methods to analyze data derived from the electronic medical record).

Also, future work should consider low frequency topics, as the analysis of the literature was focused on identifying major themes. An example of this was the displaying of high frequency words found in the title or abstract through the use of word clouds. However, it may also be useful to identify and display low frequency topics as they may present new opportunities for research as at the time of original publication the innovation may not have been further adopted due to constraints in the technology. Examples of these include concepts such as 'virtual reality' which contained five references in the dataset from 1998 to 2004 and 'augmented reality' which had one reference in 2001. With new consumer devices like the Oculus Rift and Microsoft's Kinect and HoloLens emerging, it may be useful to re-examine this earlier work.

Another key area for future work is the examination of the interaction between the user and various technologies for displaying and interacting with the visualizations. The Synthesis reference manager (involving word clouds) and Carrot2 (involving topic clustering and cluster mapping visualization) seem to be best utilized by individual users as they provide the ability to directly manipulate the application software to explore the raw literature references allowing the individual user to focus directly on his/her interests and objectives. On the other hand, the information dashboard displayed with the 39 inch 4K monitor allowed more information to be displayed due to the increase resolution and supporting small group discussion amongst two to three users due to its increased size. Further, the use of the large visualization wall with its increased resolution has the ability to aid larger group discussion. This allows broad patterns to be visualized, but at a resolution that permits drill down into details without losing sight of the initial visualized context. We anticipate that this type of technology will be increasingly used in the future with its potentially wider availability and reduced costs.

There are several limitations of this study. First, the visualizations presented in this study are technology-based and designed around the use of high-resolution screens for their optimal presentation. Furthermore, they require some degree of interaction with the user to aid in the displaying of the information. This can be seen with the word cloud contained within the Synthesis application with the dynamic slider for increasing or decreasing the number of words displayed, and the Carrot2 application with topic clustering where the parameters could be adjusted for increasing or decreasing the size of the groups. Additionally, the dashboards with the publication time line allow additional information about the literature reference to be displayed, but only when the user clicks on that reference. While the 2016 edition of the Yearbook is the 25th, for our analysis we had only 24 editions available since the 2016 edition was still in preparation at the time this paper was written.

Conclusions

Medical informatics is a vast and expanding area. As new technologies have been developed and entered into the health sector, they have created even more research, adoption, and evaluation. These changes create increasing challenges in the synthesis and conceptualization of information in the

area of medical informatics. Visualization approaches for aiding the understanding of literature is an active area of research, and like medical informatics, it is constantly evolving as new technologies, both hardware and software, are developed. This paper explores several ways of how the medical informatics literature could be visualized in both the development of aggregated summarized information, as well as, the actual presentation of the references. However, it is a challenging problem to summarize an area as diverse as medical informatics over a period of nearly 25 years. Future research into the visualization and understanding of the academic literature would be of benefit.

Competing Interests
DWY is a co-founder of Synthesis Research Inc. which owns the intellectual property for the Synthesis software application. HTT and EPM declare they have no competing interests.

Author's contributions
DWY contributed to the study design, software development, theme identification, data analysis, writing of the manuscript, and overall coordination of this study. HTT and EPM contributed equally to the study design, data analysis, and writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Dr. Daniel J Dutton, Mr. Kevin Mackie, Dr. Douglas R Hamilton, Ms. Susan Powelson, and Dr. John Brosz for their comments and feedback on this project.

References

1. Haux R, Kulikowski C, editors. IMIA Yearbook of Medical Informatics. Digitial Libraries and Medicine. Stuttgart, Germany: Schattauer Verlagsgesellschaft mbH; 2001.
2. Jaulent M, Lehmann C, Séroussi B, editors. IMIA Yearbook of Medical Informatics. Patient-Centered Care Coordination ed. Stuttgart, Germany: Schattauer GmbH; 2015.
3. Seroussi B, Jaulent MC, Lehmann CU. Health information technology challenges to support patient-centered care coordination. Yearb Med Inform 2015;10(1):8-10.
4. Deshazo JP, Lavallie DL, Wolf FM. Publication trends in the medical informatics literature: 20 years of "medical informatics" in MeSH. BMC Med Inform Decis Mak 2009;9:7-6947-9-7.
5. Mihalas GI. Evolution of trends in european medical informatics. Acta Informatica Medica 2014;22(1):37.
6. Synnestvedt MB, Chen C, Holmes JH. Visual exploration of landmarks and trends in the medical informatics literature. AMIA Annu Symp Proc 2005:1129.
7. Synnestvedt MB, Chen C, Holmes JH. CiteSpace II: Visualization and knowledge discovery in bibliographic databases. AMIA Annu Symp Proc 2005:724-8.
8. Douglas SM, Montelione GT, Gerstein M. PubNet: A flexible system for visualizing literature derived networks. Genome Biol 2005;6(9):R80.
9. Yergens D, Ray J, Doig C. KSv2: Application for enhancing scoping and systematic reviews. AMIA Annu Symp Proc 2012.
10. Yergens DW, Dutton DJ, Patten SB. An overview of the statistical methods reported by studies using the canadian community health survey. BMC Med Res Methodol 2014;14:15-2288-14-15.
11. Gill D, Griffin A. Good medical practice: What are we trying to say? textual analysis using tag clouds. Med Educ 2010;44(3).
12. Osinski S, Weiss D. Carrot2: An open source framework for search results clustering (poster). 26th European Conference on Information Retrieval, Sunderland, UK; 2004.
13. Osinski S, Stefanowski J, Weiss D. Lingo: Search results clustering algorithm based on singular value decomposition. Advances in Soft Computing, Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM´04 Conference, Zakopane, Poland; 2004:359-68.
14. Yergens D, Minty E, Doig C. Visualization of publication timelines using 4K monitors. American Medical Informatics Association (AMIA) 2014 Annual Symposium. Washington, DC.
15. Bostock M. D3.js - data-driven documents. http://d3js.org/. Updated October, 2015.
16. Massachusetts Institute of Technology and Contributors. Timeline - Web widget for visualizing temporal data. http://www.simile-widgets.org/timeline/. Updated 2009.
17. University of Calgary. Visualization studio. http://library.ucalgary.ca/viz. Updated October, 2015.

Correspondence to:
Dean W. Yergens
Department of Community Health Sciences
University of Calgary
Calgary, Alberta
Canada
E-mail: dyergens@ucalgary.ca