Using Bibliometrics-aided Retrieval to Delineate the Field of Cardiovascular Research

Diane Gal[1], Karin Sipido[1] and Wolfgang Glänzel[2]

{diane.gal, karin.sipido}@med.kuleuven.be , wolfgang.glanzel@kuleuven.be
[1]Department of Cardiovascular Sciences, KULeuven, 3000 Leuven (Belgium)
[2]ECOOM and Dept. MSI, KU Leuven, 3000 Leuven (Belgium) & Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
A hybrid search strategy, using lexical and citation based methods, is presented in this paper as a robust method to delineate the broad field of cardiovascular research. Overall, this study aims to provide scientifically reliable and accurate data driven evidence about cardiovascular research by establishing a dataset of published research in this field. A workflow is presented that outlines the methods carried out to establish a core dataset based on a core set of journals, to identify and use search terms to detect a broader dataset, and then to apply measures of similarities between the citations of these two datasets to ensure relevance of the final dataset. The final core set of journals established comprises of 120 unique journals covered in Thomson Reuters *Web of Science Core Collection* (WoS) database including a total of 320,647 documents from 1991 to 2013. The search terms utilised include 107 cardio-specific terms that initially identify 1.8 million unique documents when searching the title, abstract and keywords. Upon application of the citation-based similarity measures the final combined dataset consists of 845,071 publications. Overall, establishing a relevant dataset of cardiovascular research means placing a greater emphasis on having a precise dataset, reducing recall in the process.

Conference Topic
Methods and techniques

Introduction
Experts in the cardiovascular field are concerned that there is a decline in quality and innovation in cardiovascular research and that fragmentation of this broad field is leading to loss of cross-pollination and missed opportunities for translation of research from bench to bedside. In this context we have launched a project to examine cardiovascular research output over a 23 year period to provide rigorous and reliable scientific information about cardiovascular research activities. The findings of this project are expected to serve as a complement to expert opinion and previously published studies (Huffman et al., 2013; Jones, Cambrosio, & Mogoutov, 2011; Sipido et al., 2009; van Eck, Waltman, van Raan, Klautz, & Peul, 2013; Yu, Shao, He, & Duan, 2013), to provide scientifically reliable and accurate data driven evidence about cardiovascular research.
The objectives of the project are to:

- Characterise the size, growth, topics and visibility of research outputs over 23 years;
- Analyse the geographical distribution of research outputs and its evolution;
- Visualise and analyse research collaboration; and
- Identify emerging topics in cardiovascular research.

To gain a comprehensive view of research in this field a broad scope and definition has been applied to include papers published in scientific journals from basic, clinical and epidemiological studies related to the cardiovascular system, including the heart, the blood vessels and/or the pericardium. The main source of data is the *Web of Science Core Collection*. The purpose of this paper is to describe the methods utilised, and the roadmap set, to establish a dataset of published research undertaken in the cardiovascular field.

Methods

Hybrid search strategies for subject delineation, previously described and published (Bolaños-Pizarro, Thijs, & Glänzel, 2010; Glänzel, Janssens, & Thijs, 2009; Zitt & Bassecoulard, 2006), have been adapted to establish a dataset of cardiovascular research. This includes (1) establishing a core dataset based on a core set of journals and core search terms, (2) identifying a broader dataset of publications through the use of search terms, and then, (3) applying measures of similarities by citations between the documents in these datasets to select a final dataset with acceptable precision and recall. A workflow/roadmap was developed to outline the main steps taken to establish the dataset, as can be seen in Figure 1.

Core Journal Dataset

All data have been retrieved from Thomson Reuters Web of Science Core Collection. The core set of journals was selected through expert review of the scope/aim of all 183 journals included in the 'Cardiac & Cardiovascular Systems' and the 'Peripheral Vascular Disease' Web of Science Categories. The scope/aim for each journal was obtained through online web-based searches. Using an online survey tool, two experts reviewed the title and scope/aim of each journal to assess the relevance of the journal and indicate whether they had experience with each journal (e.g. reading, editing, reviewing, submitting a document for publication). Journals that were assessed by at least 1 expert as being a core cardiovascular journal – defined as a journal publishing greater than 90% of its articles, reviews, letters and notes on the cardiovascular domain – were included in the core journal dataset. Disagreements between the experts were reviewed by the project team. Journals were excluded from the core dataset only when the expert excluding the journal was the only one that had previous experience with the journal. The final dataset was obtained by identifying all articles, letters, notes and reviews published journals that are covered in the 1991–2013 volumes of the WoS database.

Search Terms Datasets

A number of sources were reviewed to identify relevant cardiovascular-specific search terms, including:

- Medical Subject Headings (MeSH)
- International Classification of Diseases (ICD)-10
- Cochrane Hypertension/Heart/Peripheral Vascular Disease Groups/Systematic Reviews
- Cardioscape project taxonomy (European Society of Cardiology, 2014)
- Recent published research (Bolaños-Pizarro et al., 2010; Huffman et al., 2013; Jones et al., 2011; van Eck et al., 2013)

Subsequently, a group of eight topic experts representing a mix of clinical scientists, basic scientists and epidemiologists were invited to review the combined list of 105 search terms to assess their relevance in identifying as broad a range of cardiovascular research publications as possible. All search terms were included where at least half of the reviewers agreed that they were relevant search terms to include in the search strategy.

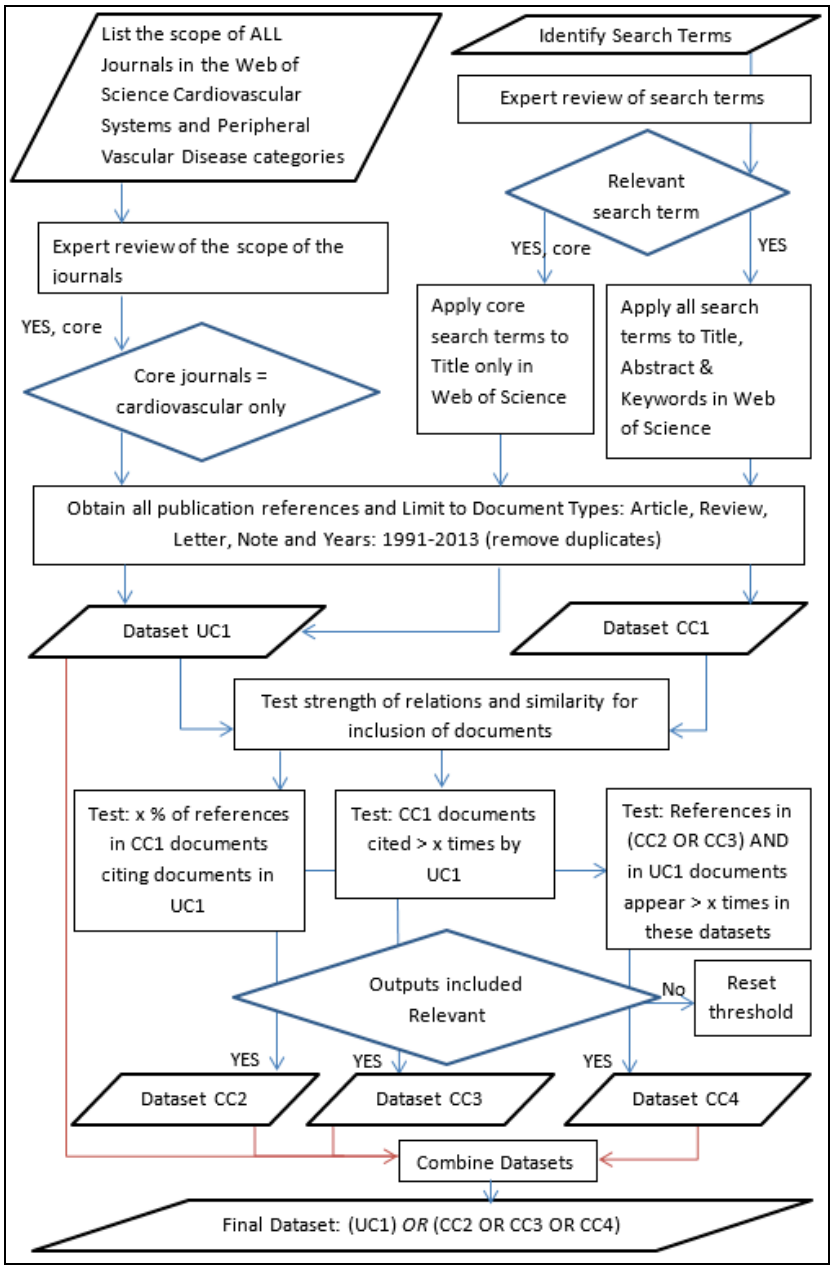

Figure 1. Workflow of field delineation of Cardiovascular Research

In addition, experts were asked to suggest any potentially missing search terms. New search terms suggested and disagreements were reviewed by the project team. The broad search terms dataset was obtained by applying the full search strategy to the complete Web of Science database, to identify all articles, letters, notes and reviews published between 1991 and 2013. To add to the core journal dataset, highly cardiovascular specific or core search terms were selected that when searched in the title would identify core cardiovascular publications.

Similarity Measures and Thresholds

For the extension of the core dataset, i.e., the seed of relevant literature, we followed an algorithm using a logical combination of *unconditional* and *conditional* criteria (Glänzel, 2014). In the present project we have linked literature retrieved based on conditional criteria (the broad search terms set) to the set of surely relevant documents (the core journals and core search terms set), using citation-based similarities. In particular, three measures of similarity

between the core dataset and the broad search terms dataset were utilised: a) the share of references of broad search terms documents that cite the core documents, b) the number of references of the core documents that cite the broad search terms documents and c) the number of shared references between the core dataset and the restricted search terms dataset. The thresholds for each measure were set following iterative testing, whereby a low threshold was first applied and a random sample of the titles and abstracts of 500 documents was reviewed for relevance to the cardiovascular field. The threshold was altered until the sample contained a high precision and the level of noise (peripheral and irrelevant documents) was reduced to an acceptable level, defined as a 5% level of noise. To confirm the relevance of the documents identified, the random samples considered to have acceptable thresholds were reviewed by one topic expert.

Findings

Core Dataset

After expert review, 120 journals were included as core journals. The two expert reviewers agreed on the exclusion of 61 journals and disagreed on the inclusion of 39 journals (21% of all 183 journals), of these only two journals were excluded as the expert who had experience with the journal was the one that excluded it. For the remaining 37 journals, they were included since both experts had previous experience for three journals and neither expert had experience for 34 journals. The final core journal documents therefore consist of 320,647 articles, letters, notes and reviews from 1991 to 2013. Thirteen of the search terms, identified below, were considered to be highly cardiovascular specific. The core search terms when searched only in the title, added 141,676 documents to the core journal documents, resulting in a core dataset of 462,323 documents. Review of this dataset confirmed that it provides a precise sample of cardiovascular-specific documents for this study.

Broad Search Terms Dataset

After expert review by 6 topic experts and the project team, 107 search terms were included in the final search strategy. Of the original 105 terms reviewed, three search terms were removed since more than half of the experts suggesting to remove them. A total of 22 unique terms were also suggested by three of the topic experts. The project team assessed and included four of these new terms. Then one additional term was added to the search strategy to include this term with and without its common prefix. The final broad search terms dataset consists of 1,656,278 unique articles, letters, notes and reviews from 1991 to 2013 where the search terms could be identified in the abstract, keywords or title. All documents in the core dataset were removed from this broad search term dataset.

A comparison of all documents obtained by searching the abstract, keywords and title is presented in Figure 2.

As a validation of the search strategy and selection of core journals, when the search strategy was applied to the 120 core journals, 95% of all core journal dataset documents were identified by the search terms.

Similarity Measures and Thresholds

An initial test was undertaken to limit the search terms dataset by removing all documents that had no links with the core journal documents. A total of 228,000 documents had no links meaning they did not cite the core journal set, they were not cited by the core journal set *and* they did not have any common references with the core journal set. This reduced the search terms set to less than 1.6 million documents, however upon review of random samples it was

clear that stronger measures of similarity would be needed to further restrict the search terms dataset to include the most relevant documents in the final dataset.

Iterative testing and review of random samples led to the selection of a combined dataset where at least 12% of the references in the broad search documents cited documents in the core dataset or where the broad search documents where cited greater than 4 times by the core documents. For this chosen dataset, no more than 10% of the random samples were considered not relevant or peripheral to the cardiovascular field. Documents from the third measure of similarity using bibliographic coupling was not included in the final dataset since it was not possible to achieve less than a 10% noise level through iterative testing and review of random samples. The final restricted broad search terms dataset consists of 382,748 unique articles, letters, notes and reviews from 1991 to 2013.

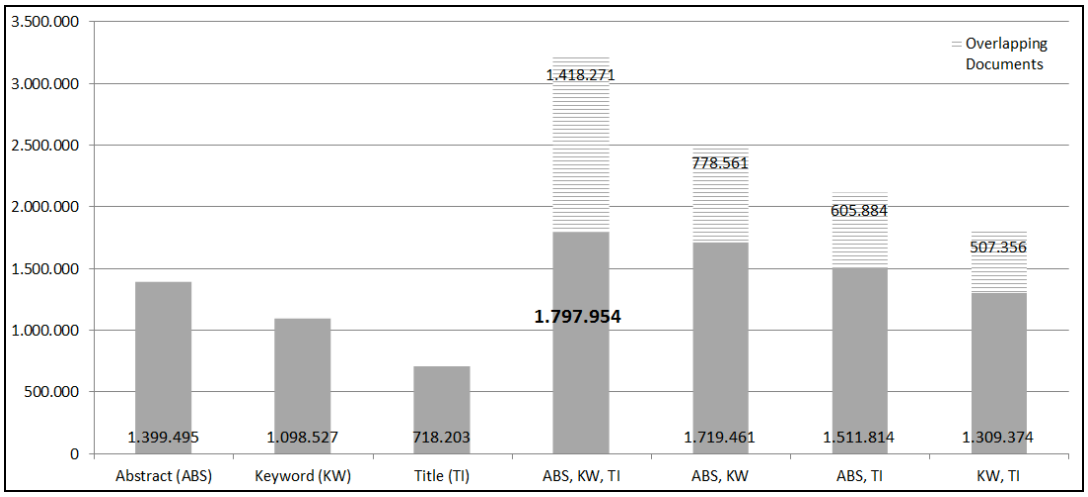

Figure 2. Number of documents identified when searching 107 search terms in Abstracts, Keywords and Titles [Data sourced from Thomson Reuters Web of Science Core Collection].

Final Combined Dataset

Combined, the core and restricted datasets create a final dataset of 845,071 unique documents from the cardiovascular field. Overall, the combined dataset has a 4.5% noise level (estimated).

Discussion

Only one previously published bibliometric study of cardiovascular research used a hybrid search strategy to establish its dataset (Bolaños-Pizarro et al., 2010). However, due to the broad scope of this study, which aims to include all types of research – from basic to clinical research, a broader list of cardio-specific search terms was created. Attention was also placed on ensuring that the search terms selected could identify cardiovascular research over the long time period of the study, as well as, enable the identification of new and emerging fields in cardiovascular research. The 107 search terms greatly increases the recall of documents, though this also means that a greater amount of noise was present in the broad search terms dataset. Hence, the importance of utilising measures of similarity between the two datasets to restrict the broad search terms dataset to include only the most relevant documents. This was done through testing various thresholds of citation-based similarities, as the final step of this robust method to delineate complex fields of research. Including both directions of citation-based similarities (ie. documents from core journals dataset citing documents in search terms dataset and vice versa) also ensures that the distribution of documents sampled is representative over time. The initial threshold of 5% noise was re-evaluated through testing

and due to the broad nature of the cardiovascular field a higher level of noise (10%) was considered acceptable as this includes peripheral research that has a component linked to cardiovascular research. The broad search terms dataset has been reduced to less than a quarter of initial documents identified to ensure the final dataset is as precise as possible and can be considered a representative sample of cardiovascular research over the 23 year period.

Conclusions

Bibliometrics-aided retrieval is a robust method to delineate the field of cardiovascular research. Through using this method, a representative dataset of cardiovascular research was established irrespective of changes in the field, such as vocabulary used, over the time-frame of this study. Overall, establishing a relevant dataset of cardiovascular research means placing a greater emphasis on having a precise dataset, reducing recall in the process.

Acknowledgments

Thank you to Bart Thijs for his input into the study methods.

References

Bolaños-Pizarro, M., Thijs, B., & Glänzel, W. (2010). Cardiovascular research in Spain. A comparative scientometric study. *Scientometrics*, *85*(2), 509–526. doi:10.1007/s11192-009-0155-2

European Society of Cardiology. (2014). *CardioScape: A survey of the European cardiovascular research landscape* (p. 52). Retrieved June 2, 2015 from: http://www.cardioscape.eu/static_file/CardioScape/PNO%20report/CardioScape_Summary%20Report_30092014.pdf

Glänzel, W. (2014). Bibliometrics-aided retrieval – where information retrieval meets scientometrics. *Scientometrics*. doi:10.1007/s11192-014-1480-7

Glänzel, W., Janssens, F., & Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics*, *79*(1), 109–129. doi:10.1007/s11192-009-0407-1

Huffman, M. D., Baldridge, A., Bloomfield, G. S., Colantonio, L. D., Prabhakaran, P., Ajay, V.S., Lewison, G., & Prabhakaran, D. (2013). Global cardiovascular research output, citations, and collaborations: A time-trend, bibliometric analysis (1999–2008). *PLoS ONE*, *8*(12), e83440. doi:10.1371/journal.pone.0083440

Jones, D. S., Cambrosio, A., & Mogoutov, A. (2011). Detection and characterization of translational research in cancer and cardiovascular medicine. *Journal of Translational Medicine*, *9*(1), 57. doi:10.1186/1479-5876-9-57

Sipido, K. R., Tedgui, A., Kristensen, S. D., Pasterkamp, G., Schunkert, H., Wehling, M., Dambrauskaite, V. (2009). Identifying needs and opportunities for advancing translational research in cardiovascular disease. *Cardiovascular Research*, *83*(3), 425–435. doi:10.1093/cvr/cvp165

Van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLOS ONE*, *8*(4). doi:10.1371/journal.pone.0062395

Yu, Q., Shao, H., He, P., & Duan, Z. (2013). World scientific collaboration in coronary heart disease research. *International Journal of Cardiology*, *167*(3), 631–639. doi:10.1016/j.ijcard.2012.09.134

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, *42*(6), 1513–1531. doi:10.1016/j.ipm.2006.03.016