Thematic atlas of Italian oncological research: the analysis of public IRCCS

Corrado Cuccurullo, Luca D'Aniello, Maria Spano

1. Introduction

This paper has been developed in the frame of the research project "V:ALERE 2019" focused on Italian public-owned Academic Medical Centers (AMCs - that is 16 public AMCs as "Aziende Ospedaliere Universitarie", 9 public AMCs as "Ex Policlinici Universitari a gestione diretta", 21 public-owned "Istituti di Ricovero e Cura a Carattere Scientifico" (IRCCS) (Ministry of Health - http://www.salute.gov.it/, 2018)). These institutions have a triple mission: research, teaching, and care, having an enormous impact on society and the nation's health.

The main aim of the project is to provide new evidences and proposals to support and advise Italian public AMCs in their quest to address their challenges.

In recent years, there is increasing recognition of the potential value of research evidence as one of the many factors considered by policymakers and practitioners. Even more, in the case of medical science, the analysis of research and its impact is indispensable, in light of its implications for public health.

The starting point for mapping a research area is to review the related scientific literature by synthesizing past research findings and then, effectively use the existing knowledge base and advanced lines of future researches. In this sense, bibliometrics becomes useful, by introducing a systematic, transparent, and reproducible review process based on the statistical measurement of science, scientists, or scientific activity (Cuccurullo *et al.*, 2016). Many research areas use bibliometric methods to explore the impact of their field, the impact of a set of researchers, the impact of a particular paper, journals taken as a reference by researchers, the input knowledge, research gaps, trends, and future opportunities (Zaho, 2010). Performance analysis and science mapping (Noyons *et al.*, 1999) are the two main bibliometric approaches for investigating a research area.

In this work, we focus on science mapping as it allows identifying and displaying themes and trends with a synchronic (Callon *et al.*, 1983) or a diachronic perspective (Cobo *et al.*, 2011). By means of science mapping techniques, namely the term co-occurrence networks, and strategic/thematic maps, we aim at providing a data visualization of strategic positioning of the different Italian public AMCs in terms of their research positioning.

In particular, we identify the research-front of different AMCs and then, we visualize them in a joint representation, useful for comparing their main research themes and at the same time their different specializations, by considering also their evolution during the years.

Mapping the dynamic positioning of Italian medical research at various levels (i.e. national, regional, AMCs type, AMC) will provide a conceptual framework for policymakers and managers to understand and manage the problems of the AMCs (e.g. appropriate funding mechanisms for financing the triple-mission). Moreover, this tool could be useful for the institutions themselves to direct their research efforts towards increasingly innovative fronts taking into account the general landscape and at the same time exploiting this information to establish collaborations with other AMCs dealing with the same research topics.

Here, the effectiveness of our strategy is showed by considering the scientific production of the last 20 years of IRCCSs specialized in the oncology research.

Corrado Cuccurullo, University of Campania Luigi Vanvitelli, Italy, corrado.cuccurullo@unicampania.it, 0000-0002-7401-8575
Luca D'Aniello, University of Campania Luigi Vanvitelli, Italy, lucadaniello94@gmail.com, 0000-0003-1019-9212
Maria Spano, University of Naples Federico II, Italy, maria.spano@unina.it, 0000-0002-3103-2342

2. Data and methodology

IRCCSs are Italian healthcare organizations of relevant national interest that drive clinical assistance in strong relation to research activities. Their mission is the continuous upgrade of healthcare. The IRCCS title is granted by the Italian Ministry of Health to a very limited number of institutes throughout the nation, and their activities are federally regulated by Legislative Decree 288/2003. They are committed to being a benchmark for the whole public health system for both the quality of patient care and the innovation skills in the field of the organization. The activity of IRCCSs relates to well-defined research areas whether they received recognition for a single subject (monothematic IRCCS) or for multiple integrated biomedical areas (polythematic IRCCS).

Among the 21 public IRCCSs in Italy, we considered the nine institutions specialized in the oncology research area (6 monothematic and 3 polythematic IRCCSs).

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was used for the selection process of the publications (Liberati *et al.*, 2009). We retrieved on Web of Science (WoS) indexing database – launched by the Institute for Scientific Information (ISI) and now maintained by Clarivate Analytics – all the publications from January 2000 to December 2019. To identify the publications related to each IRCCS, we searched by full name, part of the organization name's or by its commonly known abbreviation from the Organizations – Enhanced List available on WoS (e.g. "IRCCS FND MILANO" for the Fondazione IRCCS Istituto Nazionale Tumori Milano). We limit our search by document type and selected only Articles, Proceedings Papers, Review Articles, and Book Chapters in the English language. The records were exported into PlainText format.

Starting from our final collection, we loaded the data and converted it into R data frame using bibliometrix, an open-source tool for quantitative research in scientometrics and bibliometrics that includes all the main methods for performance analysis and science mapping (Aria and Cuccurullo, 2017).

In this preprocessing phase, for the polythematic IRCCSs (*Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Istituto Nazionale Tumori Regina Elena (IRE), IRCCS Ospedale Policlinico San Martino*) we considered only the publications dealing with oncological topics, by filtering the records with respect to the metadata "Research Areas" (SC) included in WoS.

In order to consider the publications that have a major impact in the field of oncological research, we calculated the normalized citation score (NCS), one of the most frequently used field-normalized indicators (Bornmann and Haunschild, 2016). It has been calculated by dividing the citation count of a focal paper by the average citation count of the papers published in the same field (and publication year). The normalization procedure is based on all articles published within one year (and must be repeated for publications from other years).

The citation count of the article is divided by the average number of citations in the field of the article, yielding the normalized citation score for the paper. The overall normalized citation impact of each IRCCS can be analyzed on the basis of the mean value over the publication set. This results in the mean NCS (MNCS) for the paper set. In the end, following the percentile approach, we performed our analysis only on the publications with an MNCS greater than 75% (the top 25% publications).

To map the conceptual structure of each IRCCS we conducted two related analyses: a term co-occurrence network analysis and a strategic or thematic map. The combined use of these techniques allows us to illustrate: how terms relate to each other, the main research themes within each institution, and how they develop.

The basic idea behind the term co-occurrence network analysis (Wang *et al*., 2018) is that each research field or topic can be represented as a set of terms (e.g. keywords, terms extracted from titles, or abstracts). Network representation is used to understand the themes

covered by a research field, to define which are the most important and the most recent ones; i.e., the research front. Following the network approach, we built a term co-occurrence matrix, in which each cell outside the principal diagonal contains the number of times two terms appear together in the articles (co-occur). Then, the co-occurrences among terms were normalized by the association index as proposed by Van Eck and Waltman (2009). This measure assumes values in the interval [0,1] and reflects the strength of the association among terms. Co-occurrence matrices can be seen as undirected weighted graphs; therefore, we can build a network in which each term is a node and the association between linked terms is expressed as an edge, visualizing both single terms and subsets of terms frequently co-occurring together. To detect subgroups of strongly linked terms, where each subgroup corresponds to a center of interest or to a theme of the analyzed collection, we refer to community detection algorithms (Fortunato, 2010). Here, to this end, we carried out a community detection procedure by using Louvain algorithm (Blondel *et al*., 2008).

Strategic or Thematic map (Cobo *et al*., 2011) allows plotting the themes, identified through community detection, in a bi-dimensional matrix where axes are functions of the Callon centrality and density, respectively (Callon *et al.*, 1983). Centrality can be read as the importance of the theme in the research field; while density can be read as a measure of the theme's development.

In this way, we identified the conceptual structure of each IRCCS in the three different considered time slices. Then, we standardized centrality and density values, in order to make a comparison among the research fronts of the different institutions by plotting themes in a joint map. As in classical analysis, the obtained strategic map allows defining four typologies of themes (Cahlik, 2000) according to the quadrant in which they are placed. Themes in the upper-right quadrant are known as the motor themes. They are characterized by both high centrality and density. This means that they are both developed and important for the research field. Themes in the upper-left quadrant are known as isolated themes or niche themes. They have well developed internal links (high density) but unimportant external links and so are of only limited importance for the field (low centrality). Themes in the lower-left quadrant are known as emerging or declining themes. They have both low centrality and density meaning that are weakly developed or marginal. Themes in the lower-right quadrant are known as basic and transversal themes. They are characterized by high centrality and low density. These themes are important for a research field and concern general topics transversal to the different research areas of the field. In each temporal interval, we considered the KeyWords Plus (ID) used in the different documents. The ID are words or phrases that frequently appear in the titles of an article's references but do not appear in the title of the publication itself.

Their generation is based upon a special algorithm (Garfield, 1990) that is unique to *Clarivate Analytics* databases.

3. Main results

To highlight the main research themes of oncological IRCCSs and evaluating their evolution over time, we decided to divide our timespan (2000–2019) into three-time slices.

In Table 1 the distribution of the selected publications per IRCCS in the three different periods is reported. The scientific production of institutions has increased over time. The production is constant in the three-time slices for two IRCCSs (i.e. *IRCCS Ospedale San Martino* and *Istituto Nazionale Tumori Regina Elena (IRE) IRCCS*). However, some IRCCSs produced a great number of publications in the third period with respect to the previous ones (e.g. *Istituto Tumori Bari "Giovanni Paolo II" IRCCS* and *IRCCS Centro di Riferimento Oncologico della Basilicata (CROB)*).

Table 1 Publications distribution per IRCCS in the three different time slices

Period	2000 – 2006		2007 – 2013		2014 – 2019	
Organizations	No. Of Doc	% of Doc	No. Of Doc	% of Doc	No. Of Doc	% of Doc
Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico	48	18.60	73	28.29	137	53.10
Centro di Riferimento Oncologico (CRO AVIANO)	175	28.83	186	30.64	246	40.53
Fondazione IRCCS Istituto Nazionale dei Tumori	466	22.40	753	36.20	861	41.39
IRCCS Ospedale Policlinico San Martino	147	35.25	135	32.37	135	32.37
Istituto Oncologico Veneto (IOV) IRCCS	97	13.04	338	45.43	309	41.53
Istituto Tumori Bari "Giovanni Paolo II" IRCCS	16	6.53	59	24.08	170	69.39
Istituto Nazionale Tumori IRCCS – Fondazione Pascale	147	18.85	265	33.97	368	47.18
Istituto Nazionale Tumori Regina Elena (IRE) IRCCS	121	31.35	140	36.27	125	32.38
IRCSS Centro di Riferimento Oncologico della Basilicata (CROB)	11	6.51	65	38.46	93	55.03

In Figure 1 the thematic Atlas of IRCCSs' oncological research is shown. It is worth noting that each theme, identified with the community detection, is labelled with the corresponding most frequent ID.

In the three-time slices, the production of IRCCSs is rich but they have three main themes in common: *expression*, *survival,* and *chemiotherapy*. In the first time slice (2000 – 2006) *expression* was a basic theme for many IRCCSs and only for *IRE RO* was a motor theme. The position of this theme changes over the years. In the second time slice (2007 – 2013) *expression* becomes a motor theme - high density and high centrality – for many IRCCSs and starting to shift from the upper-right quadrant to the lower-right quadrant in the third slice (2014 – 2019), consolidating its role as traditional theme - low density and high centrality. Since 2007 studies focus on s*urvival* that appeared as an emerging theme on the lower-left quadrant - low density and low centrality. In the third period, *survival* becomes a traditional theme, indicating great interest in the health care of patients by many IRCCSs.

Chemiotherapy is also a theme treated by many IRCCSs over time, always positioned to the right of the map - high centrality - in the three-time slice. From the second to the third period the *chemiotherapy* theme shift from the upper-right quadrant to the lower-right quadrant, becoming a basic theme. On the upper-left quadrant, we have observed that niche themes - low centrality and high density - have increased over time. This means that the oncological research of IRCCSs is oriented towards studies more and more specialized from 2000 to 2019.
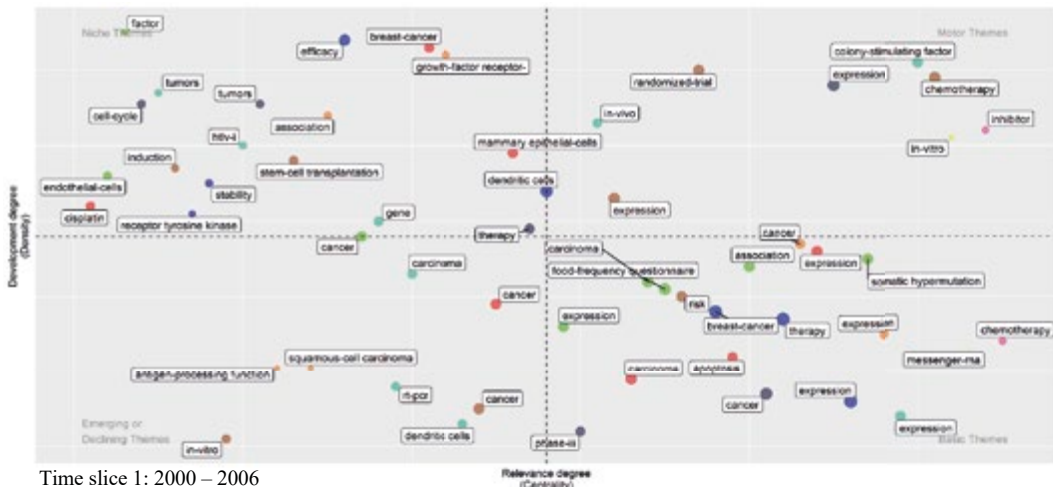
Time slice 1: 2000 – 2006

Time slice 2: 2007 – 2013
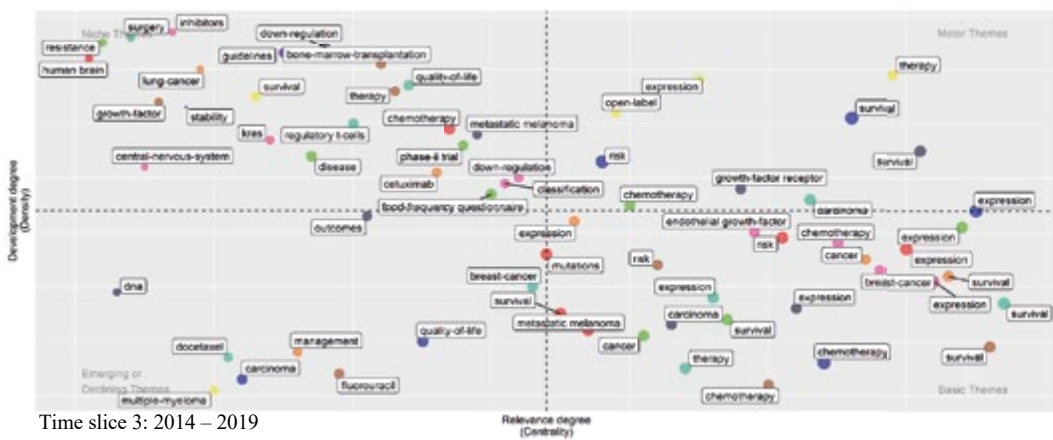
Time slice 3: 2014 – 2019

Fig. 1. Thematic Atlas of IRCCSs' oncological research

113

4. Conclusion and future developments

In this paper, we propose to jointly represent the dynamic research positioning of the different Italian public IRCCSs specialized in Oncology. These graphical representations summarize many aspects of the cancer research landscape in Italy. Obviously, the presented results are only a small part of what could be observed starting from the thematic maps.

Therefore, they are powerful decision support tools for the different agents involved in the health system. However, it is important to highlight that this approach could be used for different purposes in a more general bibliometric framework (e.g. comparison of topics covered by different sources, by different countries, or as in this case by different institutions).

On the one hand, future developments will be devoted to extending our analysis to the other Italian AMCs in order to completely mapping their research positioning; on the other hand working on the graphical representations to improve the readability of the results.

References

Aria, M., Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, **11**(4), pp. 959-975.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008. URL: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

Bornmann, L., Haunschild, R. (2016). Citation score normalized by cited references (CSNCR): The introduction of a new citation impact indicator. *Journal of Informetrics*, **10**(3), pp. 875-887.

Cahlik, T. (2000). Comparison of the maps of science. *Scientometrics*, **49**(3), pp. 373-387.

Callon, M., Courtial, J. P., Turner, W. A., Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, **22** (2), pp. 191-235.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, **5**(1), pp. 146–166.

Cuccurullo, C., Aria, M., Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains. *Scientometrics*, **108**(2), pp. 595-611.

Eck, N. J. V., Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American society for information science and technology*, **60**(8), pp. 1635-1651.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**(3), pp. 75-174.

Garfield, E. (1990). Keywords Plus®: ISI's breakthrough retrieval method. Part 1. Expanding your searching power on Current Contents on Diskette, *Current Contents*, **32**, pp. 5-9.

Liberati A., Altman D. G., Tetzlaff J., Mulrow C., Gøtzsche P. C., Ioannidis J. P. A., Clarke M., Devereaux P. J., Kleijnen J., Moher D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, **62**(10), pp. e1-e34.

Noyons, E., Moed, H., van Raan, A. (1999). Integrating research performance analysis and science mapping. *Scientometrics*, **46**(3), pp. 591-604.

Wang, H., Zhao, Y., Dang, B., Han, P., Shi, X. (2019). "Network centrality and innovation performance: the role of formal and informal institutions in emerging economies", *Journal of Business & Industrial Marketing*, **34**(6), pp. 1388-1400.

Zhao, D. (2010). Characteristics and impact of grant-funded research: a case study of the library and information science field. *Scientometrics*, **84**(2), pp. 293-306. DOI= 10.1007/s11192-010-0191-y.