How to measure productivity, quality, collaboration and innovation in Health Networks? A bibliometric study of the University of Lisbon

Carolina Bento
carolinabento98@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2021

Abstract

Scientific research plays a fundamental role in the development of society. Bibliometrics, the use of analytical and mathematical methods to study publications, has been developed to examine this research. The purpose of this thesis is to characterize the scientific production of the University of Lisbon, between 2014 and 2019, mainly in the health domain.

To accomplish this goal, the data of the publications of the institution were retrieved from the Web of Science. These data were then standardized and publications from the health domain were selected. After that, publication and citation counts, co-authorship relationships and keywords co-occurrences networks were obtained.

Productivity in the health domain grew in the periods in study from 4102 in 2014 and 2015 to 5947 in 2018 and 2019. The impact of publications of the health domain, as measured by the average number of citations per document per year, also increased regarding the first time period. Collaboration, as measured through co-authorship, has increased both institutional collaboration and international collaboration increased. Finally, innovation, as measured through the number of unique keywords and using the keywords' co-occurrence network, also presented a growing trend.

The scientific production of the University of Lisbon seems to be in ascension. More bibliometric studies should be performed to analyse this evolution, as well as to inform policy making.

Keywords: Keyword1, Keyword2, Keyword3, Keyword4, Keyword5

1. Introduction

The human era has always been characterized by progress and development. From the discovery of fire more than a million and a half years ago, followed by the invention of the wheel six millennia ago, these advancements have promoted the improvement of the quality of life for the vast majority of human beings. Food production has evolved, diseases have been eliminated or cures have been found and scientific research has contributed to the brightening of the world the human being lives in.

Nowadays, the investigation performed at universities plays a very relevant role in this progress. As such, a number of question have arisen: how is scientific research impacting our lives? What topics are being studied? How can scientific research be improved? Well, luckily, today there are thousands of tools that can be used to find an answer to these questions. The crescent computational power at the disposal of the human mind can be used to create an unprecedented understanding of science. Thus, the emergence of the multidisciplinary field "Science of Science". Quantifying or measuring scientific production, impact, collaboration and innovation may enable the discovery of unrivaled potential in all scientific fields, thus accelerating progress and the improvement of this amazing world [30].

Progress was not constructed alone, though. As science evolved, so did the way in which it was done. In the past century, the number of authors, teams, institutions, even countries, who collaborated in a single publication has been increasing [13, 17, 30]. Collaboration has been shown to increase productivity [18], impact [18, 31], funding acquisition [1, 17, 30] and even innovation [30]. However, this increase in collaboration and, consequently, its advantages are becoming more and more concentrated in the most prestigious institutions [13]. There is an homophily context in which the highest ranking institutions end up collaborating with each other and benefiting the most of this new reality contributing to crescent disparities between institutions [30]. It is therefore of extreme importance for a university to be familiar with its collaboration network and understand how to improve it as not to be left behind in the pursuit of science.

A powerful ally in the pursuit of characterizing science, and collaboration within this area, is the field of bibliometrics. Bibliometrics consists in the application of statistical and mathematical methods to assess several characteristics of scientific research. The first time someone performed a systematic acquisition of bibliographic data was in 1906 when Cattell, an American scientist and editor of *Science* between 1895 to 1944, launched the *American Men of Science*, which gathered information about influential scientists in the USA [10]. Several scientometric and bibliometric studies have been performed since then, but only with the creation of the Science Citation Index did this domain see a surge in its development. Nowadays, bibliometrics is widely used to quantify and characterize the scientific research of individual researchers, research groups, institutions, journals and even countries. These data and analysis should not be the sole information in which research managers or policy makers base their decisions on. However, allied with other methods, such as peer review, and a profound self-awareness of its own limitations, this field can provide the information needed to make reforms and improve the scientific endeavor.

Additionally, bibliometrics plays an even more prominent role in health research. This area takes advantage of developments in so many other domains that more and more data, more and more publications should be accounted for. Bibliometrics makes it easier for fields to be mapped, evolution traced and breakthroughs analysed, thus accelerating progress in the health domain [15, 16, 26]. This work aims to contribute to bibliometric literature in Portugal through the development and application of methods that characterize scientific production and collaboration of the publications of the University of Lisbon (UL), specially in the health domain.

2. Background

2.1. Bibliometrics

In order to make use of bibliometrics to assess scientific production, there are some fundamental concepts and assumptions that one has to have in mind. Firstly, the number of publications is a proxy for scientific production. Besides that, the number of citations is a proxy for the impact of a publication. Finally, co-publication is a proxy for collaboration. The first idea is based on the fact that, especially in the natural sciences and in the health domain, a publication is both a vector for information and a way to coin an idea as one's own and as such it is a means to convey scientific findings [5, 9, 25]. The use of citations as a proxy for impact – impact, not quality – is grounded on the notion that science is built from blocks, i.e. new knowledge is constructed from previously acquired knowledge, and that peers give credit to that knowledge and recognize valuable research through citations [21, 25, 28]. Finally, according to Katz and Martin [14], even though co-authorship can only be seen as a partial indicator of collaboration, it yields many advantages, as it is invariant and easily verifiable and the fact that it is not costly, while being a practical indicator.

Research output can be measured through publication counts, i.e. publication counts can be a measure of productivity [22]. Productivity may be influenced by several factors, from the individual characteristics of each researcher, to the characteristics of institutions and departments, funding and even collaboration [2, 12, 18, 19]. Lee and Bozeman defend that there is a positive relationship between the acquiring of funding and scientific fields with scientific productivity of individual researchers [18]. The fact that researchers are able to capture funding also influences productivity positively is also defended by Jacob and Lefgren [12]. Furthermore, the departmental and organizational context also influences the productivity of researchers [2]. Finally, collaboration may also be a factor that strongly influences productivity. The number of authors participating in publications seem to influence the productivity of individual researchers [18].

Besides research output, one ought to measure research impact. It is important to highlight that a higher impact does not necessarily mean that the work is of higher quality, as it may have received more attention due to other factors [25]. Having that in mind, citation counts can be used to assess attention and consequently impact [9, 21]. It is important to understand what are the reasons that may contribute to a document acquiring more or less citations. The capabilities researchers and institutions have to communicate their findings is of extreme importance to the acquisition of citations [21]. Besides that, collaboration, specially heterogeneous collaboration, i.e. the collaboration between different institutions, and international collaboration have a positive influence on impact [8].

Bibliometrics can also be used to perform science mapping, which may be extremely useful to inform decisions in research policy. According to Moral-Muñoz *et al.* [9], this method is "dedicated to showing the structural and dynamic aspects of a research field, and how it evolves through time". In order to operate such mapping, bibliometric networks are built and the basic principles of such networks are as follows: these networks are constituted by nodes, actors that can depict different units of analysis, and edges, that connect those actors and depend on the kind of network generated[9, 27]; edges are usually weighted and this weight depends both on the data and on the kind of network generated[27]; actors are placed according to relatedness, as such actors that are closer together are more related than actors further apart [28].

The focus of science mapping, like it was communicated until now, is the development of networks. Therefore science mapping is characterized by a high degree of interdisciplinary [6], with the participation of domains such as graph theory, Social Network Analysis (SNA) or even statistics [9] and the very important participation of computer science [6]. These bibliographic networks enable the recognition of fields of study as well as their visualization [6].

These networks may have different units of analysis, such as publications, individual researchers, organizations, countries, journals and words [32]. Besides that, each of these units of analysis can be "used to compile a specific kind of structure"[28]. One can develop a citation network, a co-citation network, a bibliographic coupling network, a co-authorship network and a co-word network.

2.2. Collaboration

One of the objectives of this thesis is to understand how the different colleges of the UL collaborate in the production of science in the health domain. Having that in mind, it is first necessary to assert what collaboration is. As many scholars have mentioned defining collaboration is no easy task and the fuzzy boundaries may be difficult to ascertain [3, 14]. According to Katz and Martin [14], "a 'research collaboration' could be defined as the working together of researchers to achieve the common goal of producing new scientific knowledge". Besides that definition, research collaboration between institutions can be defined as "a mutually beneficial and well-defined relationship entered into by two or more organisations to achieve common goals" [20]. However, as Katz and Martin report in their publication, the borders of research collaboration are hard to define, since scientists partake in scientific investigation in varying degrees and consequently, so do organizations [14].

One of the most used methods to measure research collaboration is co-authorship analysis [3, 14]. According to Subramanyam, using co-authorship to measure collaboration may be advantageous, because this method is "invariant; easily and inexpensively ascertainable; quantifiable and non-reactive"[24]. Nevertheless, as Buknova reminds [3] "not every research collaboration will necessarily lead to a publication and not all co-authorshiped papers are results of a collaborative research process". However, as Fonseca *et al* [7]. mentioned "The co-authorship of a technical document is an official statement of the involvement of two or more authors or organizations", and as such "co-authorship analysis is still widely used to understand and assess scientific collaboration patterns". Co-authorship can be used to build bibliometric networks and SNA can be applied to these networks to study them.

SNA has been widely used to study research collaboration and the structure of cognitive fields. Researchers have been applying the methods provided by SNA to bibliographic databases in order to perform citations and co-authorship analysis [11] or the analysis of keywords' co-occurrence networks [5]. In the case of SNA applied to co-authorship analysis, each node is an author, organization or country and two nodes are connected by an edge if there exists a co-authored paper between them [7]. When SNA is applied to keywords' co-occurrence networks, each node is a keyword and an edge between nodes exists if two keywords were associated to the same article [5].

2.3. Innovation

First of all, there is relevancy in comprehending how innovation is produced, how new ideas are devised and new technologies arise. Innovation is thought of a constant recombination of ideas, whether from ideas that already belonged to a certain field or of ideas that may have been fetched from other fields [4]. The first type of innovation is associated with "knowledge specialisation", and thus with "exploitative innovation", and the second one with "brokerage innovation", and consequently with "exploratory innovation" [4, 29]. Each type of innovation presents benefits and shortcomings. Knowledge specialisation may promote an increase in efficiency and efficacy in the processes performed by an institution or firm. Despite this, a sole investment in exploitative innovation, which has limits, may leave the institution behind in acquiring new knowledge or promoting recombinant growth [4]. As funding or investment both in academic institutions and firms is not limited there must be an equilibrium in the choice of which strategy to follow [4].

The knowledge already acquired by the institutions creates a knowledge network, in which the knowledge elements owned by the institution are connected based on their previous recombination [4]. The placement of knowledge elements in the knowledge network may influence the recombination opportunities that those elements may yet face[4].

Innovation is not a lonely activity and the position of different actors, who comprehend and own different knowledge elements, in the social network may influence the emergence of new recombinations between elements [29]. This influence may be positive or negative and it may foster more exploitative or exploratory innovation. For example, Wang *et al.* [29] found out that a researcher that presents a high degree centrality, i.e. has many connections, has less opportunities for exploratory innovation as they may be more influenced by external opinions .

To study these questions in the academia domain, a knowledge network must be devised. In order to do this, the co-word mapping already referred can possibly be used. Co-word mapping is used to delineate and understand the evolution of cognitive fields [23]. Cavadas used keyword co-occurence mapping to infer about the rise of new research topics in the area of marine research infrastructure [5].

3. Methodology

3.1. Compilation of Bibliometric Data

The first step to perform after understanding the research questions is the selection of the database (DB) for retrieval of the data. As Fonseca *et al.* mention, the DB should "cover a large number of academic journals and have high representation of health-related journals", "provide information on the affiliations of the authors, allowing the construction of organizational networks" and "the full name of the authors in most publications" and "allow the exportation of data in text format compatible with bibliometric analysis software"[7]. Both Scopus and the Web of Science (WoS) fulfill these criteria. However, the fact that the names of the institutions are more concise and coherent on the latter and that the latter is more appropriate for aggregate level analysis, WoS was the chosen DB to retrieve all the articles from the UL. Besides that, to perform the retrieval of the articles in the health domain PubMed was used.

Having in mind the objectives devised for this work, publications of the UL for a sufficiently broad time period had to be retrieved. The time period considered for the evaluation was between 2014 and 2019. The evaluation was performed jointly for the period of 2014 and 2015, then for the period of 2016 and 2017 and finally for the period of 2018 and 2019 and the query for the retrieval of the data was prepared taking that into account. Besides that, in order to retrieve the publications of all the colleges of the UL the Organization Enhanced searching tool.

3.2. Attribution of publications to the health domain

Despite the use of PubMed to classify publications as belonging to the health domain, this method was not enough. As previously mentioned, PubMed is filled with publications from the areas of medicine and the life sciences. However, these areas are not the sole areas that contribute to developments in health research. The social sciences, technology and even the arts and humanities can strengthen our understandings of the human body, health and treatments or cures.

Besides that, WoS classifies journals and books in Subject Categories and, then again the classifications may exclude publications that promote to research in health, but are not classified as medicine or life sciences. Using only this approach would belittle institutions whose primary research area is not health and would influence the results of this study. In order to overcome this issue, a method was devised to categorize publications that are not on Pubmed as being of the health domain.

3.3. Data preprocessing

In bibliometric analysis, the cleaning of the data is of extreme importance. In this project there were three stages in this step of the process, firstly the standardization of the names of the institutions' of the UL, then the articles that had been wrongly attributed to the UL were withdrawn and finally the health publications were selected.

3.4. Analysis

After the data pre-processing, it was finally possible to obtain the results and perform the analysis. The software used to perform this analysis was R and R-studio and the most important packages used were "bibliometrix", "sna" and "ggplot2".

The data were very diverse and referred to both several units of aggregation (institutions from the UL and the UL as a whole) and to different subjects (health and all the subjects in which the UL produces science) and as such the computation of the results was performed in steps. The chosen metrics to present are described in the next paragraphs.

Firstly, to evaluate the productivity of the UL, as already mentioned, publication counts were performed. This value was presented for the UL both for health and for all subjects, and for each institution of the UL considered, in the same manner, for health and for all subjects. The variation in the number of publications was also calculated. Besides that, as not all publications are the same and contribute the same to the production of scientific knowledge, the document type of each publication was considered, both for health and for all subjects, and both for the UL and for each institution.

Furthermore, each publication was attributed to a Research Area so that the influence of different areas in the scientific research produced by the UL could be described. Publications were attributed to Research Areas based on the attributions the WoS makes of each subject category. As such, publications were split into "Arts and Humanities", "Physical Sciences", "Social Sciences" and "Technology". The research area of "Life Sciences and Biomedicine" was further divided in "Life Sciences" and "Medicine". WoS attributes journals and books to at least one subject category, therefore, publications are classified into one or more subject categories, and consequently research areas, grounded on the source they were published on.

The keywords were also studied. The number of keywords and its variation and the most frequent keywords were presented both for the health domain and for all subjects, and for the UL and each institution. These could provide an insight into the cognitive field of the areas that are researched in the UL and each institution and could also help study innovation.

To study the impact of the publications of the UL, citation counts were performed. The number of citations, as well as its distribution, the average number of citations per year, the number and percentage of non-cited papers and the variation of these metrics are presented for the health domain and for all subjects for both the UL and each institution.

Finally, regarding the part of descriptive bibliometrics, to evaluate collaboration, the number of authors, the median of number of authors per article and the percentage of single-authored articles are accounted for. Besides this, the number of collaborating institutions and variation are also presented and finally the number of collaborating countries and the top 10 of most collaborative countries in the UL collaboration network. These metrics are presented for both the health domain and all subjects and for the UL.

After the descriptive bibliometric results, there was the need to analyse the networks created by the collaboration between institutions and the co-occurrence of keywords, i.e. the social structure and cognitive structure of the research performed by the UL, respectively. These networks were only analysed for the health domain. From these networks, a number of metrics that could aid in the analysis of collaboration and innovation were computed. From the metrics proposed in the literature review, only a few deemed as the most relevant were calculated, namely, the number of nodes, the number of links, the average path length, the diameter, the density and the average degree. For the most influential nodes, the number of clusters was also studied. For the collaboration network, the degree centrality of the institutions of the UL was also computed.

When talking about collaboration, the number of nodes can provide insight related to the capture of new partnerships or loss of old ones. The number of links can aid in understanding how intra-network collaboration is increasing or decreasing. The average path length, the diameter and the density are all measures of interconnectedness inside the network. A smaller average path length and diameter indicate that the distance between nodes in the network is decreasing, i.e. gaps in collaboration are bridging, while a higher density indicates that elements are becoming more connected. The average degree indicates in how many collaboration relationships the average actor in the network participates in. Finally, clusters indicate similarity between authors, and as such two actors belonging to the same cluster in a network may indicate that they collaborate with each other very much or that they collaborate very much with the same institutions.

Regarding the keyword co-occurrence network, the number of nodes may indicate an expansion or retraction of the cognitive field and the acquisition or loss of knowledge elements. The number of links in the network can be used to understand to what extent exploitative innovation is being performed. The average path length, the diameter and the density can aid in the comprehension of how interconnected the knowledge elements are. The average degree indicates with how many other knowledge elements, the average knowledge element is paired with, consequently an increase of this metric, indicates that the average element is being recombined with more elements.

3.5. Visualization

The visualization of the networks can provide insights to both specialists and people who are not in the field, as the visualization is more intuitive than numbers. As such, a software whose primary objective is to create bibliographic networks was used, VOSviewer. VOSviewer was created by van Eck and Waltman from the Leiden University and is freely available for download. In this work, VOSviewer was used through the R-package bibliometrix.

4. Results and Assessments

4.1. Productivity

As already mentioned, the research of the UL was studied for three consecutive time periods. This description began with productivity measures, such as publication counts, type of documents published and the subject categories that were researched in the UL.

In figure , it is observed that the number of publications produced by the UL has been increasing. From the first time period to the second, there was an increase of 11.3% in the number of publications produced by the UL. From the second time period to the third, there was an increase of 6.7% in the number of publications produced by the UL. The health domain presented a more significant growth in the number of publications, from 2014 and 2015 to 2016 and 2017, there was an increase of 13.6% of the publications produced by the UL in this domain, followed by a greater growth from the second time period to the third one by 27.6%. It can also be noticed that publications classified as being of the health domain constitute more than one third of the total publications produced and indexed by the UL (34.9% in 2014/15, 35.6% in the second time period and 42.6% in the third time period).

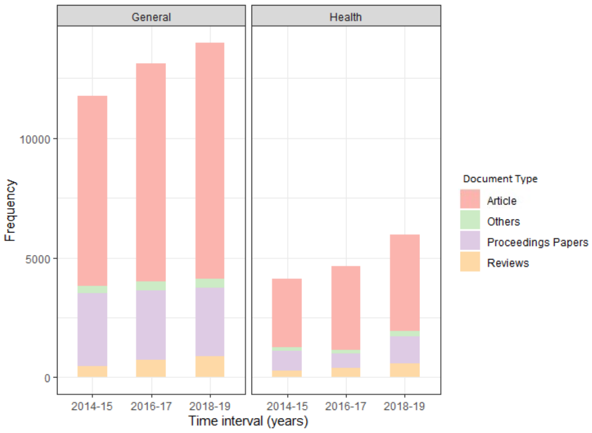

Figure 1: Number of documents produced by the UL for each of the studied time periods in this work, both in the generality of publications (left panel) and in the health domain (right panel), whose data are a subset of the left-side data. The proportion of publications that are classified as articles, proceedings papers and reviews can be found, as well as the proportion of publications that do not fit any of these categories (Others). It can be seen that publications in the health domain represent more than a third of all the publications of the UL

The document type that contributes the most to scientific production in the UL is the article constituting for the three time periods in study, and both for the generality of research produced in the UL and the health domain, more than 67% of the publications. Proceedings papers are the second most used document type to convey findings and in third place reviews can be found.

In the left panel of figure 2, it can be observed that a considerable amount (above 30%) of the publications of the University are attributed to the research area of "Technology". A significant amount of the publications (above 20%) are attributed to the field of "Physical Sciences". However, both of these fields show a slight sign of decline, while both "Life Sciences" and "Social Sciences" show signs of a steady increase in the contribution to the publications of the UL. The "Arts and Humanities" whose contribution to the publications of the UL indexed by WoS is the smallest do not show a clear pattern of growth, even though their percentage increases in both the second and third time period relatively to the first time period.

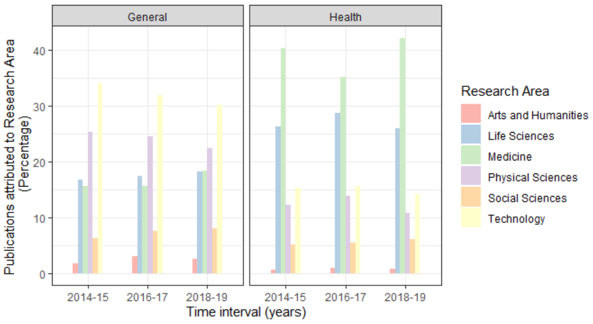

Figure 2: Percentage of publications produced by the UL attributed to each research area for each of the studied time periods in this work, both in the generality of publications (left panel) and in the health domain (right panel), whose data are a subset of the left-side data. Research areas were defined according to WoS's classification with the split of "Life Sciences and Biomedicine" in "Life Sciences" and "Medicine". The attribution of publications to each Research Area has to do with the classification of the journals or books in which they are published, and as such, a publication can figure in one or more Research Areas.

In the right panel of figure 2, the relative contribution of the distinct research areas to the research performed in the health domain can be found. As would be expected, a majority of the articles are classified in the field of "Medicine" (above 35%), followed by "Life Sciences" (above 25%). The relative contribution of the area of technology is much smaller than in the generality of publications of the UL (below 16%). The only area that shows a clear tendency of growth, even if narrow, is the area of "Social Sciences".

4.2. Impact

In figure 3, the distribution of the citations obtained by the UL in the time period in study is presented. A decrease in the number of citations can be observed both for the generality of the publications and for the publications of the health domain. Between 2014 and 2015, the median of citations for the generality of the publications was of 7 [1;19] and for the health domain it was of 11 [3;25]. For the second time period, it was of 5 [1;15], for the general domain, and of 9 [3;20], for the health domain. Finally, for the last time period, the median of the citations of the generality of the publications was of 3 [0;9] and of 5 [1;12] for the health domain. It can be observed that most publications receive few citations, while some outliers receive many. Besides that, excluding the outliers, most publications of the health domain receive more citations than the generality of publications.

When talking about citations and impact, acknowledging the percentage of documents that were not cited is necessary. In the first time period, there are 19.6% of the generality of publications that have not been cited yet and 19.4% of the publications of the health domain that are in the same situation. For the publications of the second time period, there are 20.9% of the general domain that have not been cited yet and 13.4% of the publications in the health domain that have not acquired citations so far. Fi-

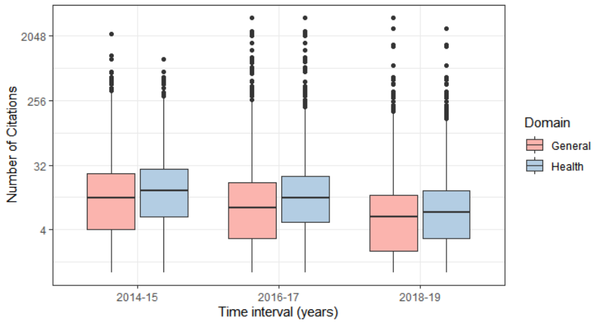

Figure 3: Distribution of the number of citations acquired by the publications of the UL in the three time periods studied in this project, both for the generality of publications and for the publications (in pink) of the health domain, which is a subset of the general publications data, (in blue), in a logarithmic scale of base 2.

nally, between 2018 and 2019, 25.9% of the generality of publications have not been cited yet and 24.4% of the publications in the health domain that have acquired zero citations so far.

4.3. Collaboration

To start drawing the picture that depicts the evolution of collaboration, one should start to look upon the individual researcher, as the collaboration process begins with this actor. In figure 4, one can study the number of authors involved in publications of the UL. It is observed that there was a significant growth in the number of authors both from the first time period to the second (31.4% for the generality of publications and 26.6% for authors that produce research in the health domain) and from the second to the third time period (28.4% for the generality of publications and 52.2% for the health domain). Besides that, in the figure we can observe that more than 50% of the researchers that collaborate in publications produced by the UL also co-author in the health domain.

In order to analyse collaboration between institutions, it is necessary to assess the amount of institutions that have participated in publications produced by the UL. In figure 5, it can be observed that the number of institutions taking part in documents published by the UL has been increasing, both for all publications and for publications in the health domain. From the first time period to the second, this growth was of 37.9%, followed by an increase of 44.5% from the second to the last studied time period, for all publications. This surge was even more substantial in the health domain, in which there was an increase of 49.6% of the institutions in the first transition between periods and of 63.8% in the second transition between periods.

Finally, collaboration is not exclusive to researchers or institutions. Scientific collaboration between countries can improve diplomatic relationships and may be a way to mitigate inequalities, as has already been mentioned in section ??. As such, to characterize the scientific production of the UL the partnerships formed with institutions from other countries should also be studied.

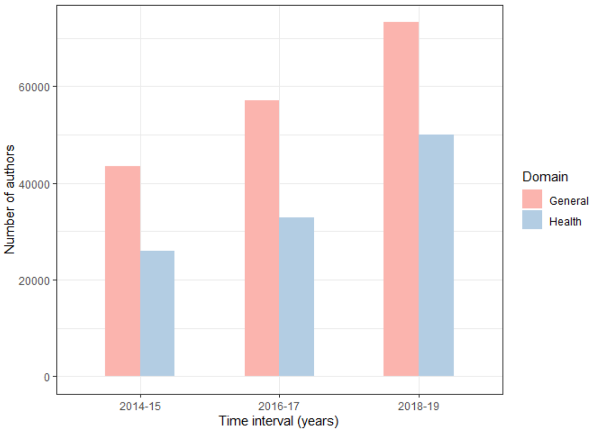

Figure 4: Number of authors that co-authored a publication of the UL for each time period in study, for both the generality of publications (in blue) and publications of the health domain, which are a subset of the all publications, (in pink)

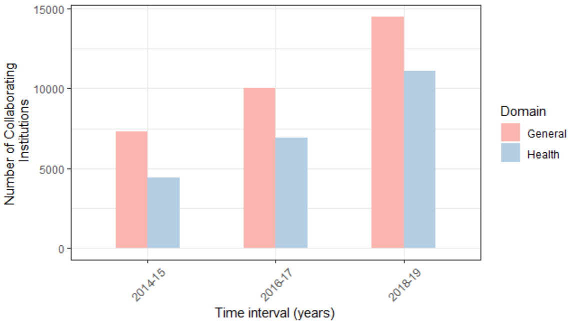

Figure 5: Number of institutions that participate in publications produced by the UL in the times period studied in this project, both for all publications (in pink) and for publications in the health domain (in blue), which are a subset of the data on the left.

Intuitively, the researchers from the UL collaborate more frequently with other researchers from Portugal. Despite this, for the generality of publications, international collaboration experienced a slight growth from the first time period to the second time period, from 75 collaborating countries to 85 countries. In the third time period, there was a maintenance of the number of collaborating countries. In the health domain, the number of countries participating in health publications of the UL increased by more than 16% in the transitions between time periods. In the first time period in study, there were 55 countries collaborating in publications of the UL, in the second time period there were 64 and in the third 74 countries were participating in the health publications produced by the

UL.

Besides that, the study of collaboration networks can provide information about the evolution of partnerships and aid in the management of collaborations. In that sense, the collaboration network, for each time period, in study of the publications in the health domain of the UL was created.

In table 1, the global metrics of the network can be found. As had already been demonstrated in the evolution of the number of institutions, the number of nodes increases, as well as the number of connections that increase by 330.6% from the first time period to the second and by 70.6% from the second to the third time period. Both the average path length and the diameter remain constant in the first transition, but decrease in the second time period. However, after a significant growth of the density of the network from 2014/15 to 2016/17, there is a relevant decline in the density from the second time period to the third. Finally, as the links increased at a faster pace than the number of nodes, a increase in the average degree can be found.

Table 1: Collaboration network's metrics for publications in the health domain produced by the UL for the three time periods studied in this project. The unit of analysis were institutions, and as such, nodes represent institutions and the links represent co-authorship of publications.

	2014/15	2016/17	2018/19
Number of Nodes	4631	6928	11351
Number of Links	520348	2240634	3822370
Average Path Length	2.36	2.36	2.21
Diameter	5	5	4
Density	0.024	0.047	0.030
Average Degree	112.36	323.42	336.74

4.4. Innovation

Creating a network of co-occurrence of keywords may also provide important information regarding the cognitive field of an institution, as well as the evolution of innovation. The keywords' networks of the three time periods can be found in the appendix and in table 2 the main metrics that characterize the networks can be found.

As already noted, there is an increase in the number of keywords used, as well as the number of links between them, which grew by 36.3%, followed by an increase of 26.5%. The average path length decreased in both transitions between time periods and the diameter, in spite of remaining constant between 2014/15 and 2016/17, declined in the third time period in analysis. Despite these contractions regarding the paths between nodes in the network, there is a shrinkage in the density of the network. As would be expected, as the number of nodes increases at a slower pace than the number of links, there is an increase in the average degree of the network. Finally, the number of clusters created by the most influential keywords presents a decreasing trend.

Table 2: Keywords' network's metrics for publications in the health domain produced by the UL for the three time periods studied in this project. The unit of analysis were the keywords authors associated to their publications, and as such, nodes represent keywords and the links represent co-occurrences of keywords.

	2014/15	2016/17	2018/19
Number of Nodes	9396	11951	14293
Number of Links	56132	76498	96792
Average Path Length	5.70	5.53	5.16
Diameter	14	14	12
Density	0.00064	0.00054	0.00046
Average Degree	5.97	6.40	6.77
Number of Clusters	55	47	40

5. Conclusions, Recommendations and Future Work

5.1. Conclusions

Scientific progress is at the heart of societal development. To build the world of tomorrow, a better, improved version of the one there is today, a more equitable, just and fair, a world that promotes equal opportunities for all, safety, peace, a world in which no child suffers from hunger or diseases that could easily be treated, a world in which each human being can advocate for their own health and well being, the role of science is going to be vital. To improve the making of science and scientific research, there needs to be a study of this area. It was in that context that this work emerged. Its goals, as already discussed, were to describe the scientific research produced by the UL, particularly in the health domain, as well as to study collaboration and innovation in this domain. It is hoped that these description and analysis may inform policy making in the UL to contribute to an amelioration of research in this institution and at a national level.

To accomplish the goals presented in the previous paragraph firstly, there was the need to come acquaintance with the world of bibliometrics and SNA, through a bibliographic review. After that, the methodology was built grounded on the literature in the area and the objectives to accomplish. It was required to retrieve the data regarding the publications that the UL produced between 2014 and 2019. The retrieval of the data was performed using the bibliographical database WoS. After the retrieval of the data, there was the need to standardize the data, particularly the affiliations' data. These data were then analysed and visualized. The main findings that were arrived using this process are summarized in the next paragraphs.

Firstly, it was found that productivity, as measured by publication counts, in the UL both for all the publications and for the publications in the health domain has increased in the studied years. This growth can be explained by several already mentioned factors, such as an increase in the human resources or funding or possibly due to the increase in collaboration.

When analysing the most frequent keywords associated to publications of the UL, it is interesting to see that health research is gaining prominence, as "Parkinson's disease" and "cancer" enter the top-10 of most used keywords. Furthermore, it can be noticed that there is a growing interest in the topic of climate change.

Regarding impact in the UL, it can be highlighted that on average publications in the health domain tend to have a higher impact than the average of all publications. Besides that, in the health domain the superior average number of citations per document per year in the second and third time period relative to the first time period in study may indicate that on the long run, publications of these time periods will have a higher impact than the ones produced in the first time period.

Concerning collaboration, there seem to be several aspects that support the hypothesis that collaboration in publications of the UL is increasing. The number of authors, institutions and countries participating in those publications. Additionally, in the health domain, the analysis of the institutional collaboration networks seem to further back the plausibility of this hypothesis.

Finally, in a very limited study of innovation, the evolution of the number of unique keywords and of the keywords co-occurrence networks may imply that innovation in the health domain in the UL is increasing.

5.2. Recommendations

As already denoted, this study is not without its limitations and there is a lot the author of this work has learned since the beginning of this work. As such, a few recommendations both to inform policy making and the future production of similar studies.

Firstly, having in mind the results and discussion of this work, it is very important to remind the reader that this study consists above all on a description of the scientific production. It is not an evaluative or comparative study neither of the UL, nor constituting institutions. Furthermore, the sole use of bibliometric data to inform policy making should be avoided and there should be a refrain from supporting the "publish or perish" science culture.

Developing from the work performed in this study, it would be relevant to carry out studies like this with more regularity, not only for the UL as a whole, but for each institution, for its departments, research groups and even individual researchers, to aid them in improving their scientific process. For this, the standardization of the affiliations and researchers' names in the publications produced is of utmost importance. Additionally, a manual to ground future bibliometric studies of the UL should be developed.

Finally, due to the impact that collaboration may have on science productivity, impact and innovation it may be relevant to foster and encourage researchers from the UL to attend inter-institutional meetings of researchers who investigate in the health domain.

5.3. Future Work

The shortcomings of the methodology used have already been presented, the future work is largely based on those shortcomings and on aspects whose analysis was not possible to perform in this study. Firstly, it would be important to improve the method to retrieve publications of the health domain, in order to obtain more accurate results. Furthermore, with the aid of data or computer science, the methods to standardize the affiliations' names should be made more rigorous. Besides that, in a future study, there should also be a pre-processing of the keywords to depict a more clear picture of knowledge domains and areas of study. Additionally, a way to overcome the biases that arose in the social sciences and the arts and humanities should be found.

The description of the collaboration and keywords co-occurrence networks is incomplete. Finding the metrics of individual nodes is of importance to understand how the network evolves and which actors promote communication the most. As such, a future study should be performed to evaluate betweenness and closeness centrality.

To further deepen this analysis, there are several hypothesis that should be studied. Studying the influence of collaboration on productivity, impact and innovation of the research produced by the UL, studying how the collaboration network of researchers and institutions influences the knowledge network or cognitive structure of the fields investigated by the UL and studying what is the contribution of the cognitive field of each institution on the cognitive field of the UL are all important to understand what policies should be put into place.

Hereafter, a study of collaboration that goes beyond scientific production should be performed.

All the recommendations and suggestions made in this section can be promising and may contribute to the development of scientific research in the UL and consequently at our scale be the drop in the ocean of science that fights for a better future.

6. Acknowledgements

This document was written and made publicly available as an institutional academic requirement and as a part of the evaluation of the MSc Thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work describe herein was developed during the period of March 2021 to November 2021 at the Unidade de Epidemiologia (UEPID) in the Faculdade de Medicina da Universidade de Lisboa, FMUL, under the supervision of Dr. Paulo Nicola. This thesis was co-supervised at Instituto Superior Técnico by Prof. Mónica Oliveira.

References

[1] G. Abramo, A. C. D'Angelo, and G. Murgia. The relationship among research productivity, research collaboration, and their determinants. *Journal of Informetrics*, 11(4):1016–1030, 2017.

[2] P. D. Allison and J. S. Long. Departmental effects on scientific productivity. *American sociological review*, pages 469–478, 1990.

[3] H. Bukvova. Studying research collaboration: A literature review. 2010.

[4] G. Carnabuci and J. Bruggeman. Knowledge specialization, knowledge brokerage and the uneven growth of technology domains. *Social forces*, 88(2):607–641, 2009.

[5] A. Cavadas. Visualising the collaboration network of a european marine research infrastructure: A bibliometric and social network analysis. *U. Porto Journal of Engineering*, 6(2):98–118, 2020.

[6] C. Chen, R. Dubin, and T. Schultz. Science mapping. In *Encyclopedia of Information Science and Technology, Third Edition*, pages 4171–4184. IGI Global, 2015.

[7] B. d. P. F. e Fonseca, R. B. Sampaio, M. V. de Araújo Fonseca, and F. Zicker. Co-authorship network analysis in health research: method and potential use. *Health research policy and systems*, 14(1):1–10, 2016.

[8] M. Franceschet and A. Costantini. The effect of scholar collaboration on impact and quality of academic papers. *Journal of informetrics*, 4(4):540–553, 2010.

[9] W. Glänzel, H. F. Moed, U. Schmoch, and M. Thelwall. *Springer handbook of science and technology indicators*. Springer Nature, 2019.

[10] B. Godin. On the origins of bibliometrics. *Scientometrics*, 68(1):109–133, 2006.

[11] V. A. Haines, J. Godley, and P. Hawe. Understanding interdisciplinary collaborations as social networks. *American journal of community psychology*, 47(1-2):1–11, 2011.

[12] B. A. Jacob and L. Lefgren. The impact of research grant funding on scientific productivity. *Journal of public economics*, 95(9-10):1168–1177, 2011.

[13] B. F. Jones, S. Wuchty, and B. Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science. *science*, 322(5905):1259–1262, 2008.

[14] J. S. Katz and B. R. Martin. What is research collaboration? *Research policy*, 26(1):1–18, 1997.

[15] P. Kokol, H. Blažun Vošner, and J. Završnik. Application of bibliometrics in medicine: a historical bibliometrics analysis. *Health Information & Libraries Journal*, 38(2):125–138, 2021.

[16] P. Kokol, J. Završnik, and H. B. Vošner. Bibliographic-based identification of hot future research topics: an opportunity for hospital librarianship. *Journal of Hospital Librarianship*, 18(4):315–322, 2018.

[17] E. Leahey. From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual review of sociology*, 42:81–100, 2016.

[18] S. Lee and B. Bozeman. The impact of research collaboration on scientific productivity. *Social studies of science*, 35(5):673–702, 2005.

[19] J. S. Long and R. McGinnis. Organizational context and scientific productivity. *American sociological review*, pages 422–442, 1981.

[20] P. W. Mattessich and B. R. Monsey. *Collaboration: what makes it work. A review of research literature on factors influencing successful collaboration.* ERIC, 1992.

[21] H. F. Moed, W. Burger, J. Frankfort, and A. F. Van Raan. The use of bibliometric data for the measurement of university research performance. *Research policy*, 14(3):131–149, 1985.

[22] R. Pranckutė. Web of science (wos) and scopus: The titans of bibliographic information in today's academic world. *Publications*, 9(1):12, 2021.

[23] H.-N. Su and P.-C. Lee. Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight. *Scientometrics*, 85(1):65–79, 2010.

[24] K. Subramanyam. Bibliometric studies of research collaboration: A review. *Journal of information Science*, 6(1):33–38, 1983.

[25] C. R. Sugimoto and V. Larivière. *Measuring research: what everyone needs to know.* Oxford University Press, 2018.

[26] D. F. Thompson and C. K. Walker. A descriptive and historical review of bibliometrics with applications to medical sciences. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 35(6):551–559, 2015.

[27] N. J. Van Eck and L. Waltman. Visualizing bibliometric networks. In *Measuring scholarly impact*, pages 285–320. Springer, 2014.

[28] A. F. Van Raan. Measuring science. In *Handbook of quantitative science and technology research*, pages 19–50. Springer, 2004.

[29] C. Wang, S. Rodan, M. Fruin, and X. Xu. Knowledge networks, collaboration networks, and exploratory innovation. *Academy of Management Journal*, 57(2):484–514, 2014.

[30] D. Wang and A.-L. Barabási. *The science of science.* Cambridge University Press, 2021.

[31] S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[32] I. Zupic and T. Čater. Bibliometric methods in management and organization. *Organizational research methods*, 18(3):429–472, 2015.