QSS

an open access journal

Citation: Sjögårde, P., & Ahlgren, P. (2019). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Science Studies*, Advance publication. https://doi.org/10.1162/qss_a_00004

DOI:
https://doi.org/10.1162/qss_a_00004

Supporting Information:
https://doi.org/10.1162/qss_a_00004

Received: 15 January 2019
Accepted: 27 July 2019

Corresponding Author:
Peter Sjögårde
peter.sjogarde@ki.se

Handling Editor:
Vincent Larivière

The MIT Press

RESEARCH ARTICLE

Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties

Peter Sjögårde[1,2] and **Per Ahlgren**[3,4]

[1]University Library, Karolinska Institutet, Stockholm, Sweden
[2]Health Informatics Centre, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
[3]Department of Statistics, Uppsala University, Uppsala, Sweden
[4]KTH Library, KTH Royal Institute of Technology, Stockholm, Sweden

Keywords: algorithmic classification, article-level classification, classification system, granularity level, specialty

ABSTRACT

In this work, we build on and use the outcome of an earlier study on topic identification in an algorithmically constructed publication-level classification (ACPLC), and address the issue of how to algorithmically obtain a classification of topics (containing articles), where the classes of the classification correspond to specialties. The methodology we propose, which is similar to that used in the earlier study, uses journals and their articles to construct a baseline classification. The underlying assumption of our approach is that journals of a particular size and focus have a scope that corresponds to specialties. By measuring the similarity between (1) the baseline classification and (2) multiple classifications obtained by topic clustering and using different values of a resolution parameter, we have identified a best performing ACPLC. In two case studies, we could identify the subject foci of the specialties involved, and the subject foci of specialties were relatively easy to distinguish. Further, the class size variation regarding the best performing ACPLC is moderate, and only a small proportion of the articles belong to very small classes. For these reasons, we conclude that the proposed methodology is suitable for determining the specialty granularity level of an ACPLC.

1. INTRODUCTION

In a recent article we proposed a methodology for identification of research topics in an algorithmically constructed publication-level classification of research publications (ACPLC; Sjögårde & Ahlgren, 2018). We used a large dataset of more than 30 million publications in Web of Science to create an ACPLC, at the granularity level of topics. We consider topics as problem areas addressed by researchers, representing "an underlying semantic theme" (Yan et al., 2012), and we see topics as the lowest level of aggregation to be considered for classification of subject areas (Besselaar & Heimeriks, 2006). However, more levels of different granularity are needed for an ACPLC to be used to answer a broader range of questions. In the present study, we use a similar methodology to create a classification whose granularity corresponds to research *specialties*. In the remainder of this paper, we use the term "specialty" instead of "research specialty." In short, a specialty is a "network of researchers who tend to study the same research topics" (Morris & Van der Veer Martens, 2008). However, the specialty notion is further discussed below. In this paper we identify the publications belonging to specialties by grouping the topics obtained in the previous study.

The identification of specialties is part of a broader aim to develop a standard approach to create a large and global hierarchical ACPLC of research publications in terms of geographical uptake, coverage of subject areas, and citation databases, such as Web of Science or Scopus. An ACPLC can be used for a great variety of analytical purposes and is especially useful for recurrent analytical activities.

A classification system that groups publications into classes whose sizes correspond to specialties can be used to study the publication output of different actors within a specialty; the collaboration between actors, dynamics, emergence and decline of specialties; and the relation between specialties. Moreover, a hierarchical classification, including both classes corresponding to topics and classes corresponding to specialties, makes it possible to identify topics within a specialty and, for example, a shifting focus of a specialty. We therefore suggest that the level of specialties, together with the level of topics, should be included in a standard ACPLC, and that such an ACPLC should be hierarchical.

The purpose of this paper is to find a theoretically grounded, practically applicable, and useful granularity level of an ACPLC with respect to specialties. To determine the granularity of specialties, a baseline classification is constructed. A set of journals is identified and used to create a baseline classification. ACPLCs with different granularities, constructed by the use of different values of a resolution parameter, are then compared to the baseline classification. The classification that best fits the baseline classification is proposed to be used for bibliometric analyses of specialties. In contrast to earlier work, our aim is to create a classification of publications that can be used to identify all specialties represented in Web of Science from 1980 onwards.

The remainder of this paper is structured as follows. In the next section, a short summary of our previous article on topic identification is given. The framework of the study is outlined in Section 3 and the specialty notion is discussed in Section 4. Data and methods are presented in Section 5, and Section 6 gives the results. Conclusions are given in Section 7.

2. SUMMARY OF THE SJÖGÅRDE-AHLGREN STUDY ON IDENTIFICATION OF TOPICS

To give the reader some background to the present study, in this section we summarize the earlier study on topic identification (Sjögårde & Ahlgren, 2018). In that study, we discussed how the resolution parameter given to the software Modularity Optimizer can be calibrated to obtain publication classes corresponding to the size of topics.

A set of about 31 million articles and reviews from Bibmet, KTH Royal Institute of Technology's bibliometric database, which contains Web of Science data, was used for the study. The study involved a methodology consisting of four steps. In the first step, we constructed a baseline classification (BCP_t) corresponding to topics, where BCP_t contains synthesis articles, operationalized as articles with at least 100 references. Each such article constitutes a class, and its list of cited references points to the reference articles of the class (i.e., to the members of the class). The underlying assumption of this approach is that synthesis publications in general address a topic.

In the second step of the methodology, several ACPLCs of different granularity with respect to the topic level were created by setting the resolution parameter of Modularity Optimizer to different values. Normalized direct citation values between the articles in the dataset were used, as proposed by Waltman and van Eck (2012). For the third step, classifications derived from the ACPLCs were obtained, where each derived classification constitutes a classification of the union of the classes of the baseline classification, BCP_t. Thus, the latter classification and

a given derived one have exactly the same underlying reference articles. In the fourth and final step of the methodology, the similarity between BCP_t and each of the derived classifications from the third step was quantified. For this purpose, the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) was used. We denoted the ACPLC such that its corresponding derived classification exhibited the largest ARI similarity with BCP_t by $ACPLC_t$.

With respect to the results of the study, the class size variation regarding $ACPLC_t$ turned out to be moderate, and only a small proportion of the articles belong to very small classes. Moreover, the outcomes of two case studies showed that the topics of the cases were closely associated with different classes of $ACPLC_t$, and that these classes tend to treat only one topic. We concluded that the proposed methodology is suitable to determine the topic granularity level of an ACPLC and that the ACPLC identified by this methodology is useful for bibliometric analyses.

In the present study, we use a similar methodology to identify specialties. The 230,559 classes obtained in the previous study, of which 136,939 have a size of at least 50 articles, are clustered into specialties. A baseline classification is constructed that corresponds to specialties, and a set of journals is used to create the baseline classification.

We need to point out that there is a substantial overlap between our earlier paper (Sjögårde & Ahlgren, 2018) and the present one. The reason for this is that the four-step methodology used in the earlier study, and briefly described above, is also used in the study underlying the present paper.

3. FRAMEWORK

As in the previous study, we use a network-based approach to obtain a classification of research publications (Fortunato, 2010). We use the Modularity Optimizer[1] software, created by Waltman and van Eck (2013), and the methodology put forward in Waltman and van Eck (2012). This framework has also been used by others (Klavans & Boyack, 2017a, b). The alternative modularity function is used (Traag et al., 2011), together with the SLM algorithm for modularity optimization. We acknowledge that a new algorithm for modularity optimization has been proposed (Traag et al., 2019). However, to be consistent with the previous study, we use the SLM algorithm in this study. We choose direct citation to express publication-publication relations, rather than bibliographic coupling (Kessler, 1965), cocitations (e.g., Marshakova-Shaikevich, 1973; Small, 1974), textual similarity (e.g., Ahlgren & Colliander, 2009; Boyack et al., 2011), or combined approaches (e.g., Colliander, 2015; Glänzel & Thijs, 2017). Direct citation is more efficient as it gives rise to fewer relations than the mentioned approaches, and there is empirical support that direct citations perform well in comparison with bibliographic coupling and cocitations when it comes to larger data sets (Boyack, 2017).

In Sjögårde and Ahlgren (2018), a network model with two levels of hierarchy, topics and specialties, was presented. This model comprises a logical classification: Each publication is classified into exactly one class at each level of hierarchy.[2] Moreover, all publications in a class, at a level below the top level, are classified into exactly one and the same parent class. It follows that each topic in the model belongs to exactly one specialty. In this study, in which we continue to use logical classifications, we obtain such a relation by clustering topics into

[1] http://www.ludowaltman.nl/slm/.

[2] A *logical classification* of a set of objects, O, is a set C of non-empty subsets of O such that (a) the union of the sets in C is equal to O, and (b) the sets in C are pairwise disjoint. Thus, each object in O is classified into exactly one set in C.

specialties, rather than using the alternative approach to cluster publications directly into specialties. Logical classifications have some shortcomings: Topics can be addressed by several specialties (Yan et al., 2012) or, at a higher level of aggregation, disciplines (Wen et al., 2017), phenomena not expressed by logical classifications. However, the relation between a topic and other specialties than the parent specialty, as well as relations between topics, can still be expressed and analyzed by use of the relational strengths associated with the edges in the model.

For further discussion on the general classification framework and for an explication of a model that expresses the relations between classes at different hierarchical levels in the model, we refer the reader to Sjögårde and Ahlgren (2018).

4. SPECIALTIES

Specialties have been studied since the 1960s in the field of sociology. In this literature, specialties are considered as smaller intellectual units within research disciplines (Chubin, 1976). The researchers within the same specialty communicate with each other. They possess similar competences and can engage in the same, or similar, research problems (Hagstrom, 1970). The notion of specialties is closely related to the notion of invisible colleges (Crane, 1972; Price, 1965). However, as pointed out by Morris and van der Veer Martens (2008), invisible colleges "presuppose that the researchers are in frequent informal contact with one another," which is not the case for specialties.

We use the definition of a specialty that has been given by Morris and van der Veer Martens (2008). They define a specialty as "*a self-organized network of researchers who tend to study the same research topics, attend the same conferences, read and cite each other's research papers and publish in the same journals*." Further, and in concurrence with others, we consider specialties to be the largest homogeneous units of science "in that each specialty has its own set of problems, a core of researchers, shared knowledge, a vocabulary, and literature" (Scharnhorst et al., 2012) and that they "play an important role in the creation and validation of new knowledge" (Colliander, 2014).

As early as 1974, Small and Griffith argued that publications can be clustered and that the obtained clusters may represent specialties (Small & Griffith, 1974). The single-linkage method was used by Small and Griffith to cluster 1,832 publications, which today would be considered a very small number of publications. They used their results to identify specialties. Since the 1970s, the technological advancements and the emergence of the Internet have changed the preconditions for research communication. There has also been a growth in research activity and production of research publications.

More lately, specialties have been identified and analyzed by the use of different clustering techniques (Lucio-Arias & Leydesdorff, 2009; Morris & van der Veer Martens, 2008; Scharnhorst et al., 2012). Different points of departure and different operationalizations of the specialty notion have captured different aspects of specialties. For example, clustering of publications based on citation relations and clustering of researchers based on coauthorship may result in different pictures of a specialty. The former approach identifies a set of publications and the latter a group of researchers belonging to a specialty. We attempt to capture the *publications* belonging to each specialty, rather than the *researchers* belonging to the specialty. A researcher can be part of several specialties, a property that cannot be expressed by the coauthorship approach. For this reason, we consider this approach less suitable for the identification of publications belonging to a specialty. We believe that it is preferable to base classifications constructed for the purpose of bibliometric analyses of specialties on the network

of publications, rather than on the network of researchers. Our approach makes it possible to identify the researchers within a specialty without forcing every researcher into exactly one specialty. It also makes it possible to analyze the contribution of one researcher to multiple specialties.

Kuhn (1996) estimates the number of core researchers in a specialty to be around 100. Based on Lotka's law (1926), Morris (2005) estimates the total number of researchers within a specialty to be around 1,000, and the number of publications produced by a specialty to be between 100 and 5,000. Boyack et al. (2014) regard specialties to be "ranging from roughly a hundred to a thousand articles per year." We acknowledge that the size of specialties in terms of publications may vary over time. Because the output of research publications has been growing the last decades, it is likely that the total size of specialties, in terms of number of publications, has been growing. Also, the yearly publication production of active specialties is likely to be on average larger today than 10 or 20 years ago. The size of specialties is an empirical question that we intend to shed light on in the present study.

5. DATA AND METHODS

As in Sjögårde and Ahlgren (2018), KTH Royal Institute of Technology's bibliometric database Bibmet was used for the study. Bibmet contains Web of Science publications from the publication year 1980 onwards. In the present study, we use the same set of publications as in the earlier study. We denote this set, in agreement with the earlier study, by P. P consists of 30,669,365 publications of the two document types: "Article" and "Review." In the remainder of this paper, we use the term "article" to refer to both articles and reviews.

5.1. Design of the Study

We attempt to find a granularity of an ACPLC, where the ACPLC is based on the articles in P, that corresponds to specialties. In order to identify the granularity of specialties, a baseline classification of publications (BCP) is created. The BCP is a set of journals, considered as classes, and each member of a class in BCP is a publication appearing in the class (i.e., appearing in the journal).

The BCP is compared to several ACPLCs with different granularities, where each such ACPLC is obtained by clustering the classes of $ACPLC_t$ (see Section 2), which is thereby utilized in the present study. An appropriate granularity is detected and an ACPLC is chosen, the classes of which correspond to specialties. The methodology, which has four steps and a high degree of similarity with the methodology proposed in Sjögårde and Ahlgren (2018), is described in detail in steps I to IV below and schematically illustrated in Figure 1.

I. Creation of baseline classes

We construct a baseline classification to correspond to specialties, which we denote by BCP_s. For the creation of BCP_s, a subset of journals covered by Web of Science is used. Each journal constitutes a class, and the publications appearing in the journal are the members of the class.

The reason for using journals to obtain BCP_s is that researchers within a specialty publish in and read the same journals. The new possibilities to search, retrieve and read research articles have changed the role of journals, but nevertheless many journals are still focused on specific areas of expertise and the researchers within those areas. Such journals aim to publish articles that are relevant to its audience. For example, we consider bibliometrics as a specialty within the discipline of library and information science, and the scope of the *Journal of Informetrics* as

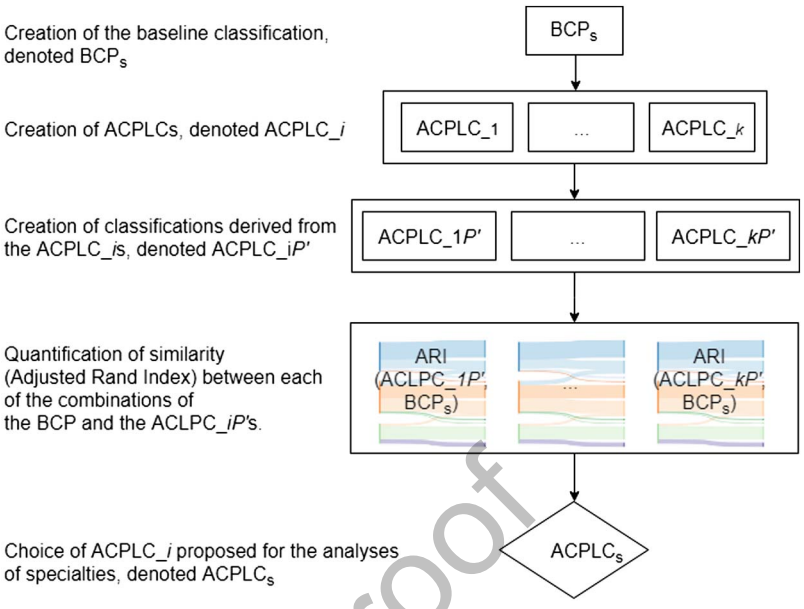

Figure 1. Illustration of the design of the study.

roughly targeting the specialty of bibliometrics. In resemblance with Bradford's law (1948), researchers within a specialty need to go to several journals to find all the relevant articles within their specialty. The boundaries of a specialty are vague and fading rather than sharp. If we consider a journal, the scope of which roughly covers a specialty, a core set of the articles in such journal is likely to be of high relevance to the core audience of the journal. The researchers that belong to this core audience can be considered as the backbone of the specialty. The rest of the articles in the journal have a fading relevance to this specialty. Some of these articles will be of higher relevance to other specialties.

When creating BCP$_s$, we attempt to delimit the set of journals to such journals that, regarding their size and scope, can be considered as proxies for specialties. As BCP$_s$ is to be used as a baseline to estimate the granularity of an ACPLC regarding specialties, the following three requirements should be addressed:

A. To be able to compare the classifications, the union of the classes in BCP$_s$ must be a subset of the union of the classes (i.e., the topics) in ACPLC$_t$.
B. Ideally, each class (journal) in BCP$_s$ should address exactly one specialty.
C. Ideally, each pair of distinct classes (journals) should address different specialties.

Now, to satisfy point A, we kept, for a given journal, only articles (i.e., publications that are of the document types "Article" or "Review") that are present in ACPLC$_t$ (i.e., having a classification at the topic level).

To deal with point B, we first delimited the publication period to five years, namely 2008–2012. By this operation, which resulted in 6,140,762 publications in 13,070 journals, the risk of including journals that, for instance, have shifted subject focus over time is lowered. In addition to dealing with point B, the choice of publications from publication years that have both incoming and outgoing citations can be assumed to have a stabilizing effect when these articles are being clustered, compared to more recent publications.

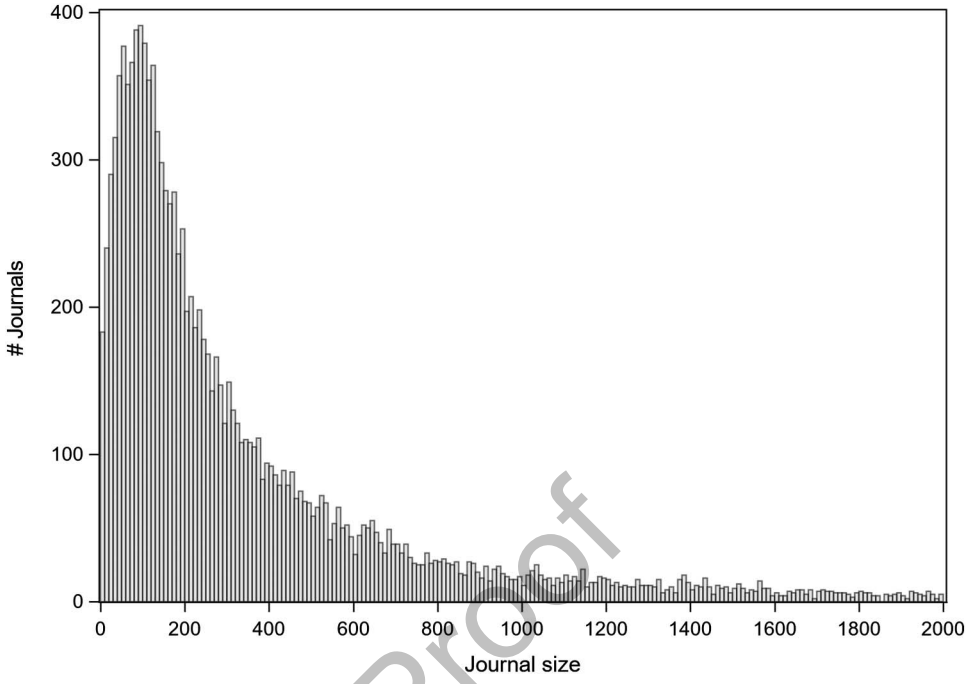

Figure 2. Number of journals per journal size for journals with 1 to 2,000 articles in 2008–2012.

We then removed all journals belonging to the Web of Science subject category "Multidisciplinary Sciences," because a journal in this category is clearly not focused on a single specialty. After this, 13,023 journals remained. Next, we considered the distribution of journals by size. Figure 2 shows the distribution limited to journals with less than or equal to 2,000 articles. A typical journal, with respect to size and modal interval as a measure of central tendency, published 90–100 articles from 2008 to 2012. By including journals between the 10th and 75th percentiles of the journal size distribution displayed in Figure 2, journals with 47–478 articles were included. With this journal size limitation, the risk to include journals addressing multiple specialties (or journals with a narrower scope than a specialty) is reduced. The limitation reduced the number of journals to 8,485.

Finally, in order to further reduce the risk of including journals addressing multiple specialties, we took journal self-citations into account. The idea is that a one-specialty journal can be assumed to cite itself to a larger extent compared to a journal that covers two or more specialties, other things held constant. In the light of this, we required, for a journal to be included in BCP_s, that the self-citation ratio (in %) should be at least 10.[3] The journal set was reduced to 1,540 journals by this procedure. Some test runs with different values of the threshold were

[3] The *self-citation ratio* (*s*) for a journal *j* is given by:

$$s_j = \frac{c_s}{r_a} \quad (1)$$

where c_s is the number of self-citations in *j*, and r_a the total number of active references in *j*. References are considered as active if they point to publications covered by the data source (Waltman et al., 2013). A reference is considered as a self-citation if the referencing publication and the referenced publication belong to the same journal.

conducted. These runs showed that lower values of the threshold reduced the maximum ARI value (cf. step IV below), which indicates that lowering the threshold value results in broader, less focused journals. The threshold was set to include as many journals as possible and to keep the ARI value reasonably high.

Some of the measures taken to satisfy point B are also relevant for satisfying point C (which states that each pair of distinct classes should address different specialties), such as the limitation to the publication years 2008–2012. With the aim to further raise the possibilities of satisfying point C, we applied bibliographic coupling between journals. If two journals had an overlap of 8% or more regarding their active cited references, they were considered as specialty overlapping.[4] This threshold was chosen after browsing the list of journal pairs sorted in descending order based on number of shared cited references. Based on journal titles, it is obvious that some journal pairs have an overlapping subject focus: for example, the two journals *Higher Education* and *Studies in Higher Education* (19% citation overlap). A threshold for the cited references overlap was chosen to include such apparent cases. In addition, test runs were conducted with different threshold values. Higher values resulted in lower maximum ARI values. For this reason, we tried to keep the threshold value as low as possible (without considering journals with nonoverlapping subject focus as specialty overlapping).

We grouped journals so that all journals that were directly or indirectly connected, by a cited reference overlap of 8% or more, were assigned the same group. For example, if journal j_1 has a cited reference overlap of ≥ 8% with journal j_2, and j_2 has a cited reference overlap of ≥ 8% with j_3, then j_1, j_2, and j_3 are assigned to the same group. Note that j_1 and j_3 are assigned to the same group even if they do not have an active reference article overlap of ≥ 8%. Each obtained group of journals was considered as addressing the same specialty. One of the journals was then randomly selected from each group. After the execution of this procedure, 967 journals remained. This number is the number of journals (classes) in BCP_s. We denote the union of the classes in BCP_s as P'.

II. Creation of ACPLCs of different granularity with respect to the specialty level

In order to obtain ACPLCs of different granularity, the first step was to measure the relatedness between the classes (topics) of $ACPLC_t$. We measured the relatedness as the average normalized direct citation value between the articles belonging to the two classes: If class C contains m articles and class C' n, the sum of the $m \times n$ normalized direct citation values between articles in C and articles in C' was divided by $m \times n$. In the second step, the generated class relatedness values were iteratively given as input to Modularity Optimizer to cluster the classes

[4] The *overlap* (y) between two journals (j_1 and j_2) is given by:

$$y = \frac{1}{2}\left(\frac{m}{A_1} + \frac{m}{A_2}\right) \quad (2)$$

where m is the number of shared cited references (i.e. cited references occurring in both j_1 and j_2), A_1 the number of cited references in j_1, and A_2 the number of cited references in j_2. The reference list of a journal was obtained by concatenating the reference lists of the articles (published year 2010) in the journal. If a reference article has been cited by more than one article in a journal, then this reference is counted multiple times for that journal. For example, if journal j_1 has four references to article a and journal j_2 has two references to article a, then journals j_1 and j_2 have two shared cited references with respect to article a. Note that we give the overlap measure threshold as a percentage in the running text.

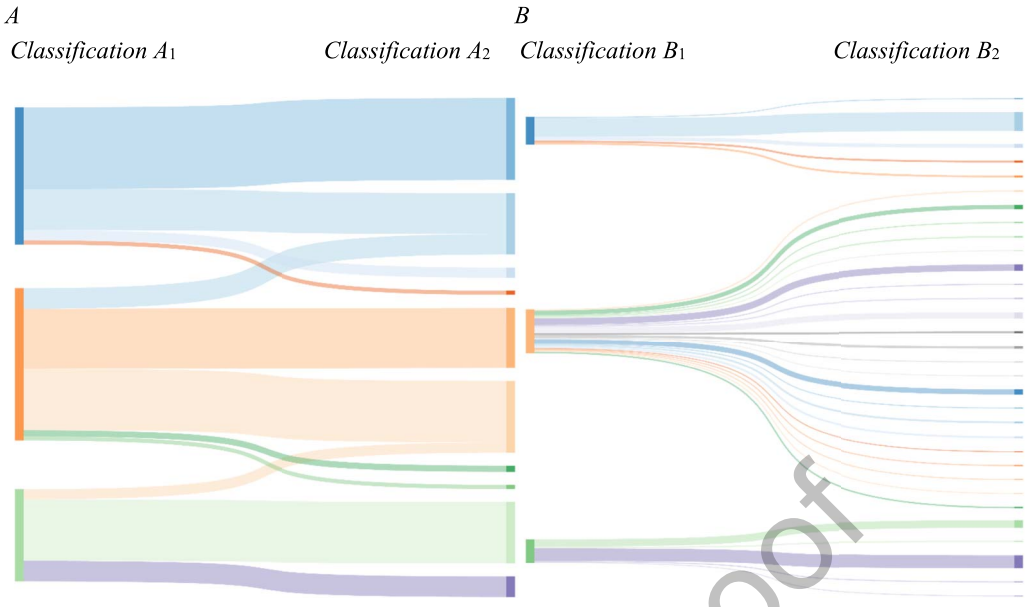

Figure 3. Two alluvial diagrams (A and B) illustrating the relation between two classifications. A shows two classifications with a high level of similarity. B shows two classifications with a low level of similarity.

of $ACPLC_t$, where the resolution parameter was set to different values in the iterations.[5] By this, ACPLCs were created for comparison of similarity with BCP^s. We denote the ACPLCs by ACPLC_1, …, ACPLC_k, where k is the number of created ACPLCs.

III. Creation of classifications derived from the ACPLCs

For each ACPLC_i ($1 \le i \le k$), a classification was derived from ACPLC_i in the following way:

(a) Each class C in ACPLC_i such that C did not contain any articles in P' was removed from ACPLC_i. Let ACPLC_i1 be the subset of ACPLC_i that resulted from the removal.
(b) For each class C in ACPLC_i1, all articles in C that did not belong to P' were removed from C. Let ACPLC_iP' be the set that resulted from these removal operations.

Clearly, the set ACPLC_iP' constitutes a classification of P' (i.e., of the union of the classes of the baseline classification BCP_s). Thus, ACPLC_iP' and BCP_s have exactly the same underlying articles. We denote the k derived classifications as ACPLC_1P', …, ACPLC_kP'. These classifications then correspond to the classifications ACPLC_1, …, ACPLC_k.

IV. Quantification of the similarity between BCP_s and the ACPLC_iP's

We attempt to optimize the granularity of an ACPLC_iP' so that it exhibits as high a similarity as possible with BCP_s. Figure 3 illustrates the relation between two classifications as an alluvial

[5] Our approach differs slightly from the approach used by Waltman and van Eck (2012). The latter approach only uses average normalized direct citation values to reassign publications (at a given hierarchical level of the classification) that belong to clusters with an insufficient number of publications. Thus, the preliminary assignment of publications to clusters, which precedes the reassignment in question, is executed without the use of average normalized direct citation values. The reason for our deviation from the Waltman–van Eck approach is that Modularity Optimizer does not directly support their approach.

diagram. Example *A* shows two classifications A_1 and A_2 with a high similarity. Example *B* shows two classifications where one of the classifications is more coarsely grained (B_1) than the other classification (B_2). The similarity between A_1 and A_2 is higher than the similarity between B_1 and B_2. If we consider B_1 as a baseline classification, then the granularity of B_2 would be too finely grained.

As in our topic identification study, we used the ARI (Hubert & Arabie, 1985) to quantify the similarity between BCP_s and an ACPLC_*iP*′. The ARI ranges from 0 to 1. It is advantageous over the original Rand Index proposed by Rand (1971), because it adjusts for chance. The ARI compares two classifications by considering pairs of items in one of the classifications and whether or not each pair is grouped into the same class in the other classification. Note that an ARI value of 1 between BCP_t and an ACPLC_*iP*′ corresponds to a situation in which these two classifications are identical. For further information on ARI, we refer the reader to Sjögårde and Ahlgren (2018).

To find the ACPLC_*iP*′ with the highest ARI similarity with BCP_s, we tested the similarity after each run of Modularity Optimizer. A first run was made with a resolution parameter value of 5E-7. This value was chosen based on previous experience and some testing. We then increased the parameter value by 5E-7. This increase resulted in a higher ARI similarity, and we therefore increased the resolution further by 5E-7 for the third run, from 1E-6 to 1.5E-6. We continued by increasing the resolution by 5E-7 in total four more times, and thus seven runs were done. The fifth run, with a resolution parameter value of 2.5E-6, gave rise to the highest ARI similarity (see Table 2 and Figure 4, Section 6).

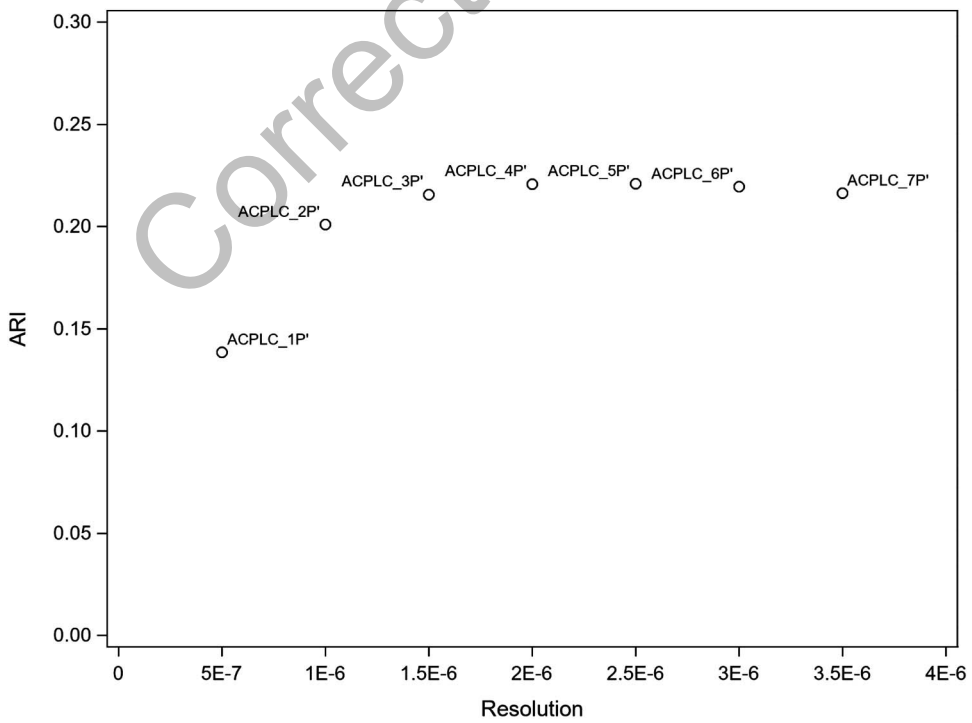

Figure 4. ARI values between ACPLC_i*P*′s and BCP_s. The vertical axis shows the ARI value and the horizontal axis shows the value of the resolution parameter used to obtain the corresponding ACPLC_*i*s. The order of ACPLC_*iP*′s corresponds to their order in Table 2.

In total BCP_s consists of 967 baseline classes. A given ACLPC_iP' consists of 202,647 articles, which is about 3.3% of the articles from the years 2008–2012 in the corresponding ACPLC_i. The ACPLC_i such that ACLPC_iP' exhibits the largest ARI similarity with BCP_s is proposed to be used for the analyses of specialties. We denote this ACPLC_i by $ACPLC_s$.

6. RESULTS AND DISCUSSION

In this section, we first deal with the selection and properties of $ACPLC_s$. Then, as in the earlier study on topic identification (Sjögårde & Ahlgren, 2018), we consider two cases. We examine the specialties of articles belonging to (1) the Web of Science subject category "Information science & Library Science," and (2) the Web of Science subject category "Medical Informatics."

6.1. Selection and Properties of $ACPLC_s$

Figure 4 shows a scatterplot of the relation between the resolution value (horizontal axis) used to obtain ACPLC_is and the ARI value (vertical axis), obtained by comparing the ACPLC_iP's with BCP_s. ACPLC_5P' has the highest ARI value. ACPLC_5P' corresponds to ACPLC_5, which we consider to be the most proper ACPLC_i with respect to granularity of specialties. In the remainder of this paper, we denote ACPLC_5 as $ACPLC_s$. However, we acknowledge that ACPLC_4P' and ACPLC_6P' have ARI values that are only slightly lower/higher than the value of ACPLC_5P'. Thus, ACPLC_4P' and ACPLC_6P' perform almost as well as ACPLC_5P'.

To get a picture of how well $ACPLC_s$ matches BCP_s, we calculated the distribution of articles in an average class in BCP_s into classes (journals) in $ACPLC_s$. This was done by first calculating the average number of classes in $ACPLC_s$ into which the articles in a class in BCP_s are distributed, an average that is equal to 50 (after rounding to the nearest integer). We then selected all 12 classes in BCP_s that were distributed into exactly 50 classes. Let the set of these classes be P_{sc}. The average number of articles in a P_{sc} class is 160.3. For each of the P_{sc} classes, we calculated the number of its articles in each of the 50 $ACPLC_s$ classes and sorted the resulting table in descending order. The $ACPLC_s$ class with the highest number of articles (i.e., the class corresponding to the first row in the table) was assigned rank 1, the second largest class (i.e., the class corresponding to the second row in the table) was assigned rank 2, etc. In this way, 12 ranked tables were obtained. Finally, averages of the number of articles by rank number, 1,…, 50, were calculated across all the 12 tables. Figure 5 shows the resulting average distribution of articles in P_{sc} (to the left) into the 50 $ACPLC_s$ classes (to the right). Ranks and average number of articles across the P_{sc} classes are shown for $ACPLC_s$.

Given that we consider the classes in $ACPLC_s$ as specialties, the distribution of journal articles in a typical BCP_s class follows a skewed distribution of specialties. About 41% of the articles in an average BCP_s class are distributed into the two most frequent specialties, and 34 specialties (classes 17 to 50) are represented by a single article (after rounding to nearest integer). Hence, a high share of the articles of the average BCP_s class is concentrated to a few of the $ACPLC_s$ classes. We therefore consider the match between $ACPLC_s$ and BCP_s as good.

$ACPLC_s$ consists of 61,805 classes, ranging from 1 to 46,078 articles. Most of the classes are small in size; however, these classes contain a small share of the total number of articles in $ACPLC_s$. For instance, classes with fewer than 500 articles contain about 1.2% of the articles in $ACPLC_s$. Figure 6 shows a histogram of the distribution of classes by class size (in terms of number of articles). In order to include classes of a substantial size in the figure, classes with fewer than 500 articles have been excluded from the figure.

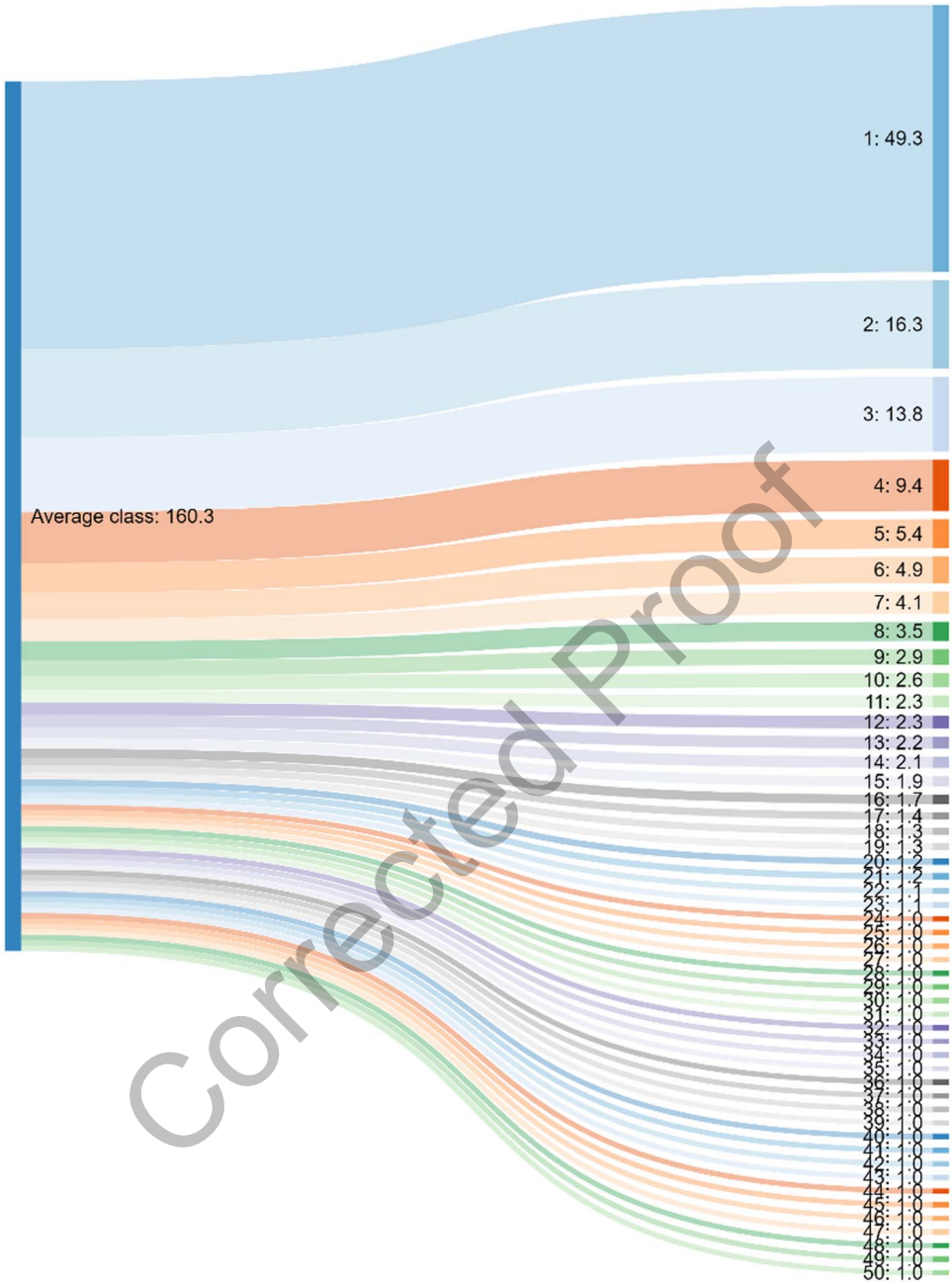

Figure 5. Alluvial diagram for an average class. The diagram shows the distribution of journal articles in BCP_s into $ACPLC_s$.[6]

Most specialties of substantial size (minimum of 500 articles) have 5 (10th percentile) to 62 (90th percentile) subordinated topics of substantial size (a minimum of 50 articles), with a mode of 6, a median of 19 and a mean of about 28 (Figure 7 and Table 1).

In Figure 8, class sizes are plotted by rank order for $ACPLC_s$ (= ACPLC_5), as well as for ACPLC_4 and ACPLC_6. A log-10 scale is used on both the vertical axis (showing class sizes by number of articles) and the horizontal axis (showing ranks). In this figure, all classes are

[6] http://sankeymatic.com/ has been used for the illustration.

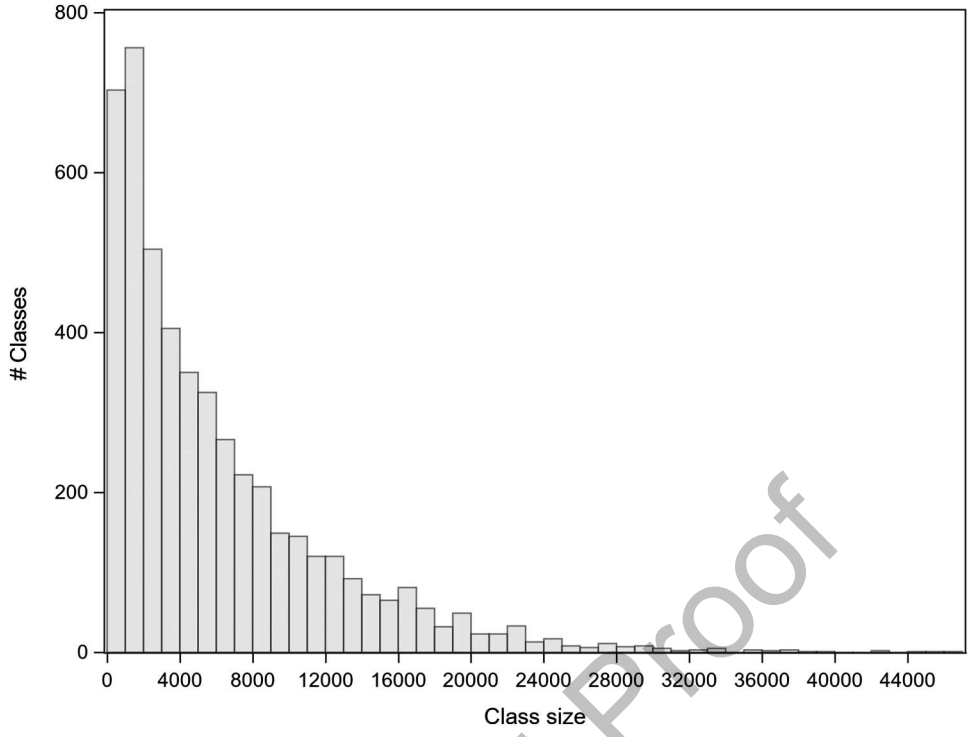

Figure 6. Histogram of number of classes by class size for $ACPLC_s$. Classes with fewer than 500 articles disregarded.

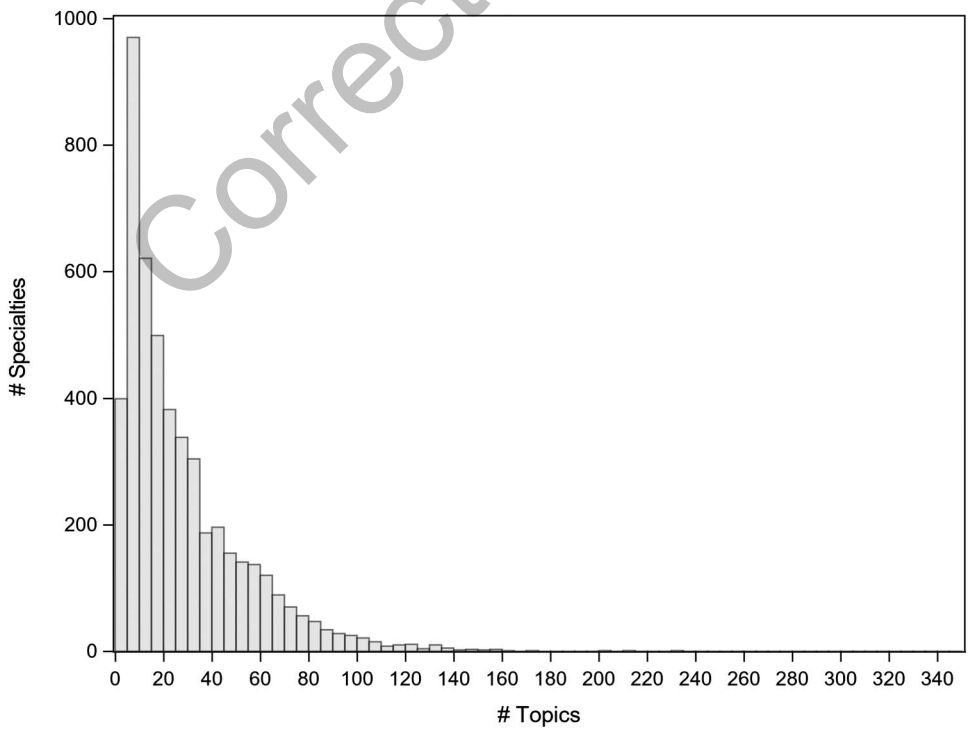

Figure 7. Histogram of number of specialties by number of subordinated topics for $ACPLC_s$. Specialties with fewer than 500 articles and topics with fewer than 50 articles disregarded.

Table 1. Distribution statistics of number of topics per specialty for $ACPLC_s$. Specialties with fewer than 500 articles and topics with fewer than 50 articles disregarded

Mean # topics per specialty	Median # topics per specialty	Mode # topics per specialty	P_{10}	P_{90}
27.6	19	6	5	62

shown, including small size classes. About 4,200 classes contain at least 1,000 articles, about 1,000 classes contain at least 10,000 articles and about 30 classes contain at least 30,000 articles. In agreement with our study on topics, the size of classes is dropping rather slowly, regardless of classification. The increasing granularity—from ACPLC_4 via $ACPLC_s$ to ACPLC_6—is reflected by, for example, corresponding, decreasing intercepts.

Figure 9 expresses the number of articles in *P* (vertical axis) that is associated with different class sizes (horizontal axis). For a randomly selected article *a*, it is most probable that the size of the specialty class in $ACPLC_s$ to which *a* belongs is 6,000–7,000 articles (cf. the highest bar of the histogram in Figure 9). Eighty percent of the articles belong to classes consisting of 2,899 (10th percentile) to 22,819 (90th percentile) articles (Table 2). The median value of $ACPLC_s$ is 10,499 and the mean 12,016. This distribution is not as skewed as the corresponding topic distribution (Sjögårde & Ahlgren, 2018, Figure 8).

The number of articles contributing to a specialty in 2015 (the most recent complete year at the time for data extraction) is between 148 and 1,597, given that we only take the mid-80% of the distribution into account (Table 3 and Figure 10). The median class size is 593. The mean number of articles per specialty class is growing approximately linearly across the 10-year

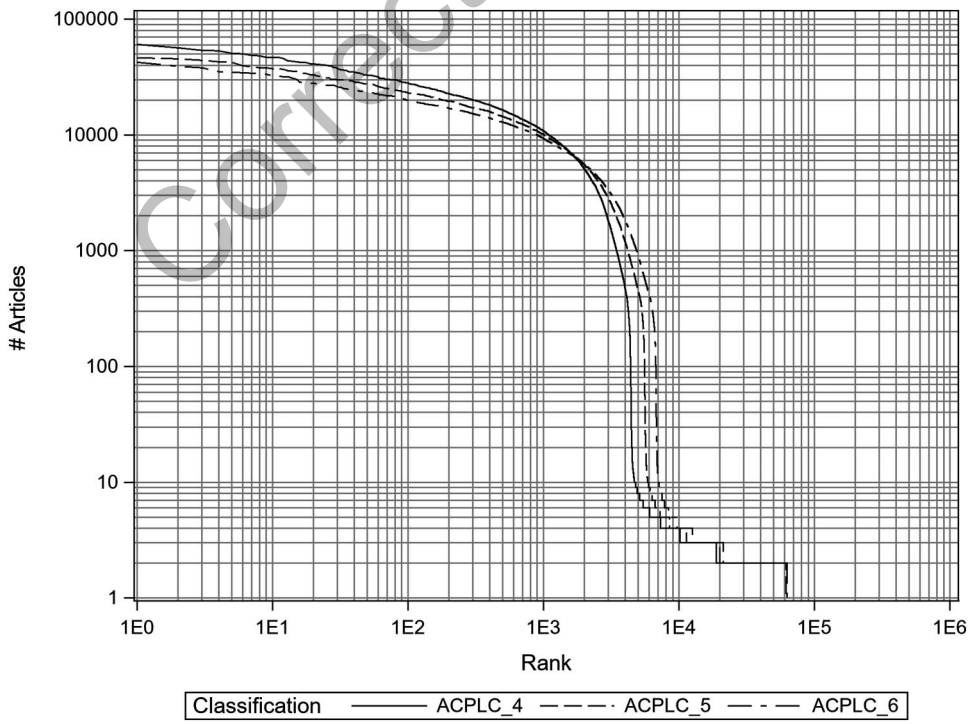

Figure 8. Distribution of number of articles by class size for three classifications. The classes in ACPLC_3, ACPLC_4 = $ACPLC_s$, and ACPLC_5 are ordered descending by size with respect to the horizontal axis. Log-10 scale used for both axes.

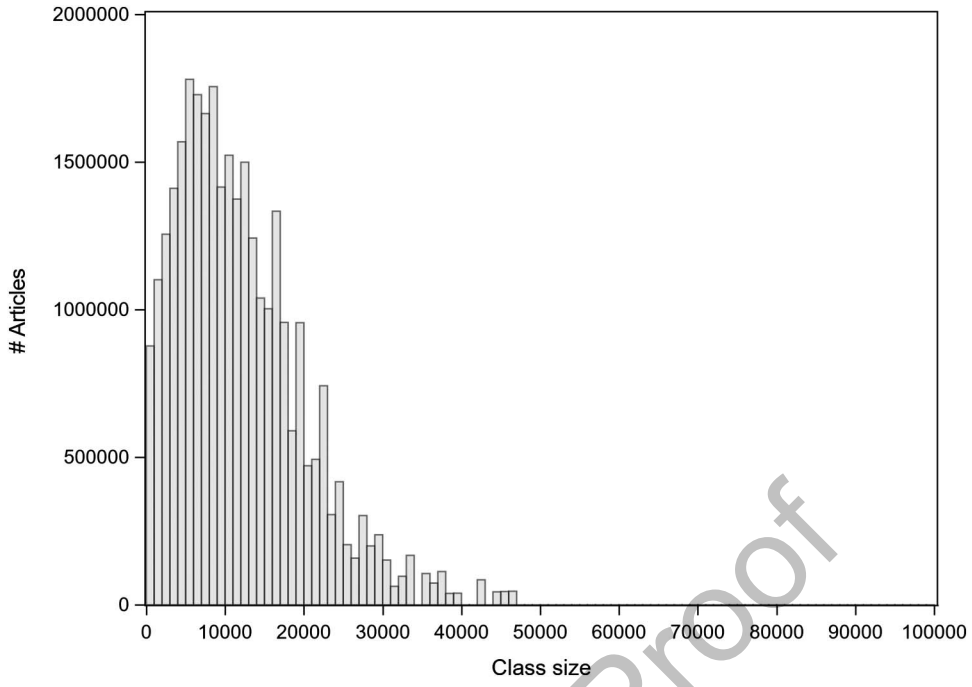

Figure 9. Histogram of number of articles by class size for $ACPLC_s$.

period (Table 3). This can be expected, considering the linear growth of research publications in Web of Science.

As mentioned in the introduction, Morris (2005) estimates the size of specialties to be between 100 and 5,000 articles (but not mentioning any time period), and Boyack et al. (2014) estimate the yearly article output of a specialty to be somewhere between 100 and 1,000 articles. The results of the present study cannot be easily compared to these figures. The

Table 2. For each ACPLC_iP', the ARI value between ACPLC_iP' and BCP_s, and the value of the resolution parameter used to obtain ACPLC_i are shown, as well as number of classes with at least 500 articles and class size distribution measures for ACPLC_i

				Weighted class size distribution measures regarding ACPLC_i (i = 1, …, 7): mean, median, 10th and 90th percentiles (denoted P_{10} and P_{90})			
Denotation	**Resolution**	**ARI value**	**# classes with # articles ≥ 500**	**Mean # articles per class**	**Median # articles per class**	P_{10}	P_{90}
ACPLC_1P'	0.0000005	0.1385	881	66,750	57,984	19,552	121,981
ACPLC_2P'	0.0000010	0.2010	1,888	31,123	27,377	8,866	59,985
ACPLC_3P'	0.0000015	0.2157	2,953	20,426	17,960	5,260	39,326
ACPLC_4P'	0.0000020	0.2208	3,969	15,228	13,145	3,765	29,509
ACPLC_5P'	0.0000025	0.2209	4,897	12,016	10,499	2,899	22,819
ACPLC_6P'	0.0000030	0.2195	5,770	9,936	8,589	2,342	18,655
ACPLC_7P'	0.0000035	0.2163	6,604	8,564	7,429	1,900	16,351

Table 3. For a 10-year period (at the time for data extraction), the table shows class size distribution measures for $ACPLC_s$

		Weighted distribution measures regarding $ACPLC_s$: mean, median, 10th and 90th percentiles (denoted P_{10} and P_{90})			
Publication year	**# Articles**	**Mean # articles per class**	**Median # articles per class**	P_{10}	P_{90}
2006	989,420	438	366	98	869
2007	1,040,026	461	384	102	918
2008	1,115,118	497	415	111	974
2009	1,166,665	525	437	114	1,028
2010	1,210,495	555	454	118	1,109
2011	1,290,309	603	484	126	1,216
2012	1,358,175	647	516	132	1,302
2013	1,435,835	705	551	140	1,434
2014	1,478,273	749	572	144	1,513
2015	1,524,010	789	593	148	1,597

estimations of Morris and Boyack et al. are rough. Morris does not mention any time period. Further, the work by Morris is rather old and the size of specialties may have increased in terms of publication output. Table 3 shows that the number of articles in Web of Science has been growing by more than 50% between 2006 and 2015. In 2015, the size of specialties ranges

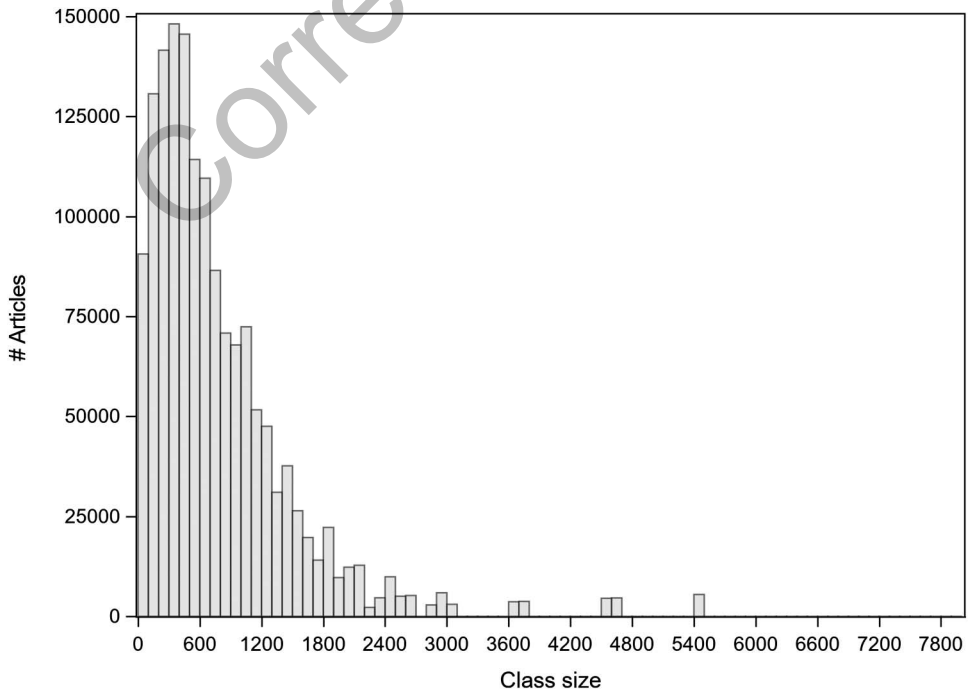

Figure 10. Histogram of number of articles by class size for the publication year 2015 and for $ACPLC_s$.

from about 150 articles (10th percentile) to 1,600 (90th percentile) articles. Thus, the size of specialties in 2015 is about 50% larger than the size estimated by Boyack et al. We regard this difference as rather small, taking into account that Boyack et al. define the next larger level (disciplines) to range from tens to hundreds of thousands of articles per year, several orders of magnitude larger than our estimation of the size of specialties.

In agreement with Morris and Boyack et al., we find it reasonable not to consider publication classes under some threshold to be regarded as specialties. One solution to the problem of small class sizes is to reassign such classes (classes below a threshold) based on their relations with larger classes (classes above or equal to the same threshold) as proposed by Waltman and van Eck (2012). However, how to set the threshold is a question that we do not address in this paper.

6.2. The Case of Information Science & Library Science

To explore how articles within the discipline of library and information science (LIS) are distributed into classes in $ACPLC_s$, we retrieved all articles in P that belong to a journal classified into the Web of Science subject category "Information Science & Library Science" and published in the period 2011–2015. In total, 16,278 articles were retrieved. Let P_{lis} be this set of articles.

For each class in $ACPLC_s$, labels were automatically created based on author keywords. Chi-square was used to quantify the relevance of author keywords in each class, and for each class, three author keywords with highest rank were concatenated to a label (for more detail see Sjögårde & Ahlgren, 2018). To distinguish the scope of each specialty, we used these labels and the labels of the topics in each class. Recall that $ACPLC_s$ is obtained by clustering the topics of $ACPLC_t$, the best performing ACPLC with respect to topic identification (Sjögårde & Ahlgren, 2018).

Table 4 shows the total number of articles in the 10 most frequent specialties and the number, and the share, of articles in a specialty that belong to P_{lis}. The top 10 specialties cover about 48% of the articles in P_{lis}. Some of the top 10 specialties are highly concentrated within the analyzed Web of Science subject category (e.g., "INFORMATION LITERACY//PUBLIC LIBRARIES//ACADEMIC LIBRARIES," 79%), whereas other specialties have a low share of its total number of articles in this category (e.g., "INNOVATION//PATENTS//OPEN INNOVATION," 7%).

The highest ranked specialty, "BIBLIOMETRICS//CITATION ANALYSIS//IMPACT FACTOR," focuses on bibliometric indicators, mapping and evaluation of research, and the analysis of scholarly communication. We acknowledge that a majority of the largest topics in this specialty are the same topics that were observed in the case study of *Journal of Informetrics* in the previous topics study (Sjögårde & Ahlgren, 2018, and Appendix 1 in this paper). The second-ranked specialty, "INFORMATION LITERACY//PUBLIC LIBRARIES//ACADEMIC LIBRARIES," focuses on library science. This category includes topics such as information literacy, knowledge organization, information practices and reference services. The specialty "INTERLENDING//DOCUMENT DELIVERY//ACADEMIC LIBRARIES" includes topics specifically related to academic libraries, such as electronic media, open access, interlending, library circulation systems and data repositories. The scopes of specialties 4, 6, 7, 8 and 10 are captured rather well by their labels, and these specialties are all clearly related to LIS. These five specialties include information retrieval, knowledge management, library and information aspects of health service and occupation as well as of innovation and patents. The specialty "ENTERPRISE RESOURCE PLANNING//ENTERPRISE RESOURCE PLANNING ERP//END USER COMPUTING" includes some topics related to LIS (e.g., IT business value, IT outsourcing, Information system planning, and Information infrastructure). The LIS relevance of "UNIVERSAL SERVICE//TELECOMMUNICATIONS//ACCESS PRICING" (rank 9) is within topics such as Internet access and Digital divide.

Table 4. Distribution of articles in the Web of Science subject category "Information Science & Library Science" into specialties, 2011–2015

Rank	Specialty	# articles in P_{lis}	Total # articles in specialty	Share of specialty in P_{lis}
1	BIBLIOMETRICS//CITATION ANALYSIS//IMPACT FACTOR	1,867	4,486	42%
2	INFORMATION LITERACY//PUBLIC LIBRARIES//ACADEMIC LIBRARIES	1,635	2,068	79%
3	INTERLENDING//DOCUMENT DELIVERY//ACADEMIC LIBRARIES	1,243	1,759	71%
4	RECOMMENDER SYSTEMS//COLLABORATIVE FILTERING//INFORMATION RETRIEVAL	564	4,965	11%
5	ELECTRONIC HEALTH RECORDS//ELECTRONIC MEDICAL RECORD//MEDICAL INFORMATICS	494	3,724	13%
6	ENTERPRISE RESOURCE PLANNING//ENTERPRISE RESOURCE PLANNING ERP//END USER COMPUTING	484	1,481	33%
7	KNOWLEDGE MANAGEMENT//KNOWLEDGE SHARING//OPEN SOURCE SOFTWARE	457	1,439	32%
8	INNOVATION//PATENTS//OPEN INNOVATION	366	5,519	7%
9	UNIVERSAL SERVICE//TELECOMMUNICATIONS// ACCESS PRICING	326	1,219	27%
10	HEALTH LITERACY//INTERNET//MHEALTH	314	4,394	7%

Appendix 1 lists the 10 topics with most publications in P_{lis} for the top 10 ranked specialties with regard to P_{lis}.

6.3. The Case of Medical Informatics (MI)

In analogy with the case of LIS, we retrieved all articles in P that belong to a Web of Science subject category, in this case "Medical Informatics," and published in the period 2011–2015, to explore how articles within this discipline are distributed into classes in $ACPLC_s$. In total, 12,516 articles were retrieved. Let P_{mi} be this set of articles.

Table 5 shows the top 10 specialties in P_{mi}, ranked by frequency. Only one specialty is highly concentrated into the "Medical Informatics" category, namely "ELECTRONIC HEALTH RECORDS//ELECTRONIC MEDICAL RECORD//MEDICAL INFORMATICS" (which is also present in the LIS case). For the rest of the top 10 specialties, 14% or less of the articles in the specialty belong to P_{mi}. This might suggest that MI is more interdisciplinary than LIS. It can also be the case that MI articles are published in broader journals, which are not classified into the "Medical Informatics" Web of Science subject category.

The largest specialty in the "Medical Informatics" category focuses on clinical decision support systems, clinical research informatics and electronic health records. The second-ranked specialty within the category, "HEALTH LITERACY//INTERNET//MHEALTH," addresses topics within mobile health such as personal health records, online health information, and online support groups. The specialty "HEALTH TECHNOLOGY ASSESSMENT//EQ 5D//PRIORITY SETTING" focuses on health technology assessment and cost effectiveness.

Table 5. Distribution of articles in the Web of Science subject category "Medical Informatics" into specialties, 2011–2015

Rank	Specialty	# articles in P_{mi}	Total # articles in specialty	Share of specialty in P_{mi}
1	ELECTRONIC HEALTH RECORDS//ELECTRONIC MEDICAL RECORD//MEDICAL INFORMATICS	1,548	3,724	42%
2	HEALTH LITERACY//INTERNET//MHEALTH	628	4,394	14%
3	HEALTH TECHNOLOGY ASSESSMENT//EQ 5D// PRIORITY SETTING	316	3,142	10%
4	ADAPTIVE DESIGN//INTERIM ANALYSIS//DOSE FINDING	297	2,094	14%
5	MISSING DATA//MULTIPLE IMPUTATION// GENERALIZED ESTIMATING EQUATIONS	288	2,530	11%
6	COMPETING RISKS//INTERVAL CENSORING// COUNTING PROCESS	286	2,278	13%
7	EVIDENCE-BASED MEDICINE//PUBLICATION BIAS//ABSTRACT	206	3,646	6%
8	PATIENT SAFETY//MEDICATION ERRORS//MEDICAL ERRORS	192	3,983	5%
9	TELEMEDICINE//TELEHEALTH//TELEPATHOLOGY	186	2,482	7%
10	CAUSAL INFERENCE//PROPENSITY SCORE// PRINCIPAL STRATIFICATION	141	1,545	9%

The remaining seven top 10 ranked specialties have the following foci: (4) clinical trial designs; (5) mathematical and statistical models and methods within the medical sciences; (6) prediction and risk models; (7) evidence-based medicine, medical epistemology, meta-analysis methods, and literature searching; (8) Patient safety (includes incident and error reporting); (9) telehealth (can be seen as a predecessor to mobile health); and (10) gene ontologies.

Appendix 2 lists the 10 topics with most publications in P_{mi} for the top 10 ranked specialties with regard to P_{mi}.

7. CONCLUSIONS

In this study we have discussed how the resolution parameter given to the Modularity Optimizer software can be calibrated to cluster topics, obtained in a previous study on topic identification (Sjögårde & Ahlgren, 2018), so that the obtained publication classes correspond to the size of specialties. A set of journals has been used as baseline for the calibration. Journals were selected based on their size and self-citation rate. The underlying assumption of our approach is that journals of a particular size and focus have a scope that corresponds to specialties. By measuring the similarity between (1) the baseline classification and (2) multiple classifications obtained by using different values of the resolution parameter, we have identified a classification, which we denote as $ACPLC_s$, whose granularity corresponds to specialties.

Some criteria for the evaluation of $ACPLC_t$, the best performing ACPLC with respect to topic identification, are the same for the evaluation of $ACPLC_s$. The differences in class sizes should not be too large and "the number of very small clusters should be minimized as much as possible" (Šubelj et al., 2016). In $ACPLC_s$, 80% of the articles belong to classes consisting of 2,899–22,819 articles. Further, 80% of the articles belong to classes with a yearly publication

rate of 98–869 articles in publication year 2006, increasing to 148–1,597 in the publication year 2015. Only 1.2% of the articles in $ACPLC_s$ belong to classes with a total number of articles less than 500. As in the previous study, the distribution follows a typical scientometric distribution, and we therefore consider the results, regarding class sizes, as satisfying.

In the present study, we have not implemented a reclassification of small classes. However, in accordance with the previous study, we consider reclassification of small classes to be desirable for practical reasons. Moreover, we think that content labeling of classes is a topic for future work.

Another criterion stated by Šubelj et al. (2016) is that classes should make intuitive sense. In addition, we stress that the focus of a specialty should be possible to identify and that two specialties should have subject foci that can be distinguished. Two case studies, in which we have identified specialties within the disciplines of LIS and MI, have been performed to evaluate these criteria. We could identify the subject foci of the specialties in these case studies, and the subject foci of the specialties have been relatively easy to distinguish. Thus, the two criteria are (approximately) satisfied in our case. Further, several of the specialties identified in the LIS case have been identified by others (Bauer et al., 2016; Blessinger & Frasier, n.d.; Figuerola et al., 2017; Janssens et al., 2006) and the same holds for several of the specialties identified in the MI case (Kim & Delen, 2018; Schuemie et al., 2009; Wang et al., 2017). However, more case studies are needed to verify the soundness of the methodology used.

The aforementioned feature of the classification approach used in this study, logical classification, which assigns each topic to exactly one speciality, has some limitations. It is clear that topics can be addressed by several specialties (or at higher level disciplines). For instance, Appendix 1 and 2 show that the topic with the label "NATURAL LANGUAGE PROCESSING// MEDICAL LANGUAGE PROCESSING//CLINICAL TEXT" is addressed by both the LIS and MI disciplines. This topic is forced into exactly one specialty, "ELECTRONIC HEALTH RECORDS//ELECTRONIC MEDICAL RECORD//MEDICAL INFORMATICS." Thus, relations between this topic and, for example, specialties within the LIS discipline are not expressed by $ACPLC_s$. However, relations between a specialty and topics within other specialties can still be analyzed using, for instance, citation relations. Nevertheless, a logical classification to some extent oversimplifies the complex structure of topic representation in research publications.

We acknowledge that direct citations perform less well than bibliographic coupling in a recent study (Waltman et al., 2019). However, a relatively low number of articles was used in the study, about 700,000 in comparison with the over 31 million articles used in this study. Moreover, a relatively short publication window (2007–2016), in comparison with the present study (1980–2016), was used. Interestingly, the study shows that an extended direct citation approach, in which direct citation relations within an extended set of publications are taken into account, performs better than an ordinary direct citation approach. Which publication-publication similarity measure to be used for the creation of an ACPLC still needs to be further investigated, however.

We recognize that there is only a small difference in performance, regarding the ARI values, between ACPLC_4P', ACPLC_5P' and ACPLC_6P'.[7] Therefore, we can only determine the granularity of specialties roughly. This is, however, not surprising, given the complex, overlapping nature of research subject areas. Nevertheless, this study sets a benchmark of the size of specialties and outlines a methodology for the calibration of ACPLCs.

[7] ACPLC_4P' had the highest ARI value in an earlier version of the manuscript. In that version, BCP_s was delimited to publications from 2010.

The combined outcome of our previous study on the classification of topics, and the present study on the classification of specialties, is a two-level hierarchical classification. We believe that such a classification comprises a valuable part of a research information system and propose that such a classification can be used for bibliometric analyses of topics and specialties.

ACKNOWLEDGMENTS

We would like to thank two anonymous reviewers for their relevant and constructive comments on an earlier version of this paper.

FUNDING INFORMATION

No funding has been received.

COMPETING INTERESTS

The authors have no competing interests.

AUTHOR CONTRIBUTIONS

Peter Sjögårde: Conceptualization; methodology; software; formal analysis; writing—original draft; writing—review & editing; visualization. Per Ahlgren: Conceptualization; methodology; formal analysis; writing—original draft; writing—review & editing.

DATA AVAILABILITY STATEMENT

The data analyzed in this manuscript is subject to copyright (by Clarivate Analytics®, Philadelphia, Pennsylvania, USA) and cannot be made available.

REFERENCES

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49–63. https://doi.org/10.1016/j.joi.2008.11.003

Bauer, J., Leydesdorff, L., & Bornmann, L. (2016). Highly cited papers in Library and Information Science (LIS): Authors, institutions, and network structures. *Journal of the Association for Information Science and Technology*, *67*(12), 3095–3100. https://doi.org/10.1002/asi.23568

Besselaar, P. van den, & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics*, *68*(3), 377–393. https://doi.org/10.1007/s11192-006-0118-9

Blessinger, K., & Frasier, M. (n.d.). *Analysis of a Decade in Library Literature: 1994–2004 | Blessinger | College & Research Libraries*. https://doi.org/10.5860/crl.68.2.155

Boyack, K. W. (2017). Investigating the effect of global data on topic detection. *Scientometrics*, *111*(2), 999–1015. https://doi.org/10.1007/s11192-017-2297-y

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management*, *32*, 147–159. https://doi.org/10.1016/j.jengtecman.2013.07.001

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R. et al. (2011). Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. *PLoS ONE*, *6*(3), e18029. https://doi.org/10.1371/journal.pone.0018029

Bradford, S. C. (1948). *Documentation*. London: Lockwood.

Chubin, D. E. (1976). State of the field: The conceptualization of scientific specialties. *Sociological Quarterly*, *17*(4), 448–476. https://doi.org/10.1111/j.1533-8525.1976.tb01715.x

Colliander, C. (2015). A novel approach to citation normalization: A similarity-based method for creating reference sets. *Journal of the Association for Information Science and Technology*, *66*(3), 489–500. https://doi.org/10.1002/asi.23193

Colliander, Cristian. (2014). *Science mapping and research evaluation: A novel methodology for creating normalized citation indicators and estimating their stability* (Doctoral thesis). Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:752675

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: University Of Chicago Press.

Figuerola, C. G., García Marco, F. J., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, *112*(3), 1507–1535. https://doi.org/10.1007/s11192-017-2432-9

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and "core documents" for the representation of clusters and topics: The astronomy dataset. *Scientometrics*, *111*(2), 1071–1087. https://doi.org/10.1007/s11192-017-2301-6

Hagstrom, W. (1970). Factors related to the use of different modes of publishing research in four scientific fields. In E. C. Nelson & K. D. Pollock (Eds.), *Communication Among Scientists and Engineers* (pp. 85–124). Lexington, MA: Heath Lexington Books.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing & Management*, *42*(6), 1614–1642. https://doi.org/10.1016/j.ipm.2006.03.025

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, *16*(3), 223–233. https://doi.org/10.1002/asi.5090160309

Kim, Y.-M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. *Health Informatics Journal*, *24*(4), 432–452. https://doi.org/10.1177/1460458216678443

Klavans, R., & Boyack, K. W. (2017a). Research portfolio analysis and topic prominence. *Journal of Informetrics*, *11*(4), 1158–1174. https://doi.org/10.1016/j.joi.2017.10.002

Klavans, R., & Boyack, K. W. (2017b). Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd edition). Chicago, IL: University of Chicago Press.

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, *16*, 317–323.

Lucio-Arias, D., & Leydesdorff, L. (2009). An indicator of research front activity: Measuring intellectual organization as uncertainty reduction in document sets. *Journal of the American Society for Information Science and Technology*, *60*(12), 2488–2498. https://doi.org/10.1002/asi.21199

Marshakova-Shaikevich, I. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy*, (6), 3–8.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, *56*(12), 1250–1273. https://doi.org/10.1002/asi.20208

Morris, S. A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology*, *42*(1), 213–295. https://doi.org/10.1002/aris.2008.1440420113

Price, D. J. de S. (1965). *Little Science, Big Science*. New York: Columbia University Press.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.2307/2284239

Scharnhorst, A., Börner, K., & Besselaar, P. (2012). *Models of Science Dynamics*. Springer: Berlin, Heidelberg.

Schuemie, M. J., Talmon, J. L., Moorman, P. W., & Kors, J. A. (2009). Mapping the domain of medical informatics. *Methods of Information in Medicine*, *48*(1), 76–83.

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies*, *4*(1), 17–40.

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLoS ONE*, *11*(4), e0154404. https://doi.org/10.1371/journal.pone.0154404

Traag, V. A., Waltman, L., & Eck, N. J. van. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233. https://doi.org/10.1038/s41598-019-41695-z

Traag, V., Dooren, P., & van Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, *84*(1), 016114. https://doi.org/10.1103/PhysRevE.84.016114

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471. https://doi.org/10.1140/epjb/e2013-40829-0

Waltman, L., van Eck, N. J., van Leeuwen, T. N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, *7*(2), 272–285. https://doi.org/10.1016/j.joi.2012.11.011

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2019). *A Principled Methodology for Comparing Relatedness Measures for Clustering Publications*. Retrieved from ArXiv:1901.06815 [Cs], http://arxiv.org/abs/1901.06815.

Wang, L., Topaz, M., Plasek, J. M., & Zhou, L. (2017). Content and trends in medical informatics publications over the past two decades. *Studies in Health Technology and Informatics*, *245*, 968–972.

Wen, B., Horlings, E., van der Zouwen, M., & van den Besselaar, P. (2017). Mapping science through bibliometric triangulation: An experimental approach applied to water research. *Journal of the Association for Information Science and Technology*, *68*(3), 724–738. https://doi.org/10.1002/asi.23696

Yan, E., Ding, Y., & Jacob, E. K. (2012). Overlaying communities and topics: An analysis on publication networks. *Scientometrics*, *90*(2), 499–513. https://doi.org/10.1007/s11192-011-0531-6

APPENDIX

Appendix 1. Topics per Specialty – LIS

SPECIALTY TOPIC	# articles in P_{lis}	# articles in topic	Share of topic in P_{lis}
BIBLIOMETRICS//CITATION ANALYSIS//IMPACT FACTOR			
FIELD NORMALIZATION//SOURCE NORMALIZATION// RESEARCH EVALUATION	193	258	75%
H INDEX//HIRSCH INDEX//G INDEX	190	317	60%
RESEARCH COLLABORATION//SCIENTIFIC COLLABORATION// CO AUTHORSHIP	125	184	68%
AUTHOR CO CITATION ANALYSIS//BIBLIOGRAPHIC COUPLING// CO CITATION ANALYSIS	83	142	58%
GOOGLE SCHOLAR//SCOPUS//WEB OF SCIENCE	79	121	65%
ALTMETRICS//MENDELEY//RESEARCHGATE	75	113	66%
OVERLAY MAP//SCIENCE OVERLAY MAPS//JOURNAL CLASSIFICATION	75	148	51%
CO AUTHORSHIP NETWORKS//SCIENTIFIC COLLABORATION// CO AUTHOR NETWORKS	70	136	51%
BOOK CITATION INDEX//SOCIAL SCIENCES AND HUMANITIES// BOOK PUBLISHERS	65	93	70%
WEBOMETRICS//WEB VISIBILITY//WEB LINKS	60	80	75%
INFORMATION LITERACY//PUBLIC LIBRARIES//ACADEMIC LIBRARIES			
INFORMATION LITERACY//INFORMATION LITERACY INSTRUCTION// LIBRARY INSTRUCTION	194	212	92%
LIBRARY 20//LIBRARIAN 2//ACADEMIC LIBRARIES	104	110	95%
KNOWLEDGE ORGANIZATION//FACETED CLASSIFICATIONS// INDEXING LANGUAGE	93	106	88%
INTERACTIVE INFORMATION RETRIEVAL//END USER SEARCHING// INFORMATION NEEDS AND USES	92	128	72%
INFORMATION PRACTICES//AIDS TALK//BARRIERS TO INFORMATION SEEKING	85	99	86%
INFORMATION SCIENCE//DIKW HIERARCHY//PROPERTIES OF DOCUMENTARY PRACTICE	81	99	82%
PUBLIC LIBRARIES//CHILDRENS INTERNET PROTECTION ACT// RURAL LIBRARIES	62	68	91%
REFERENCE SERVICES//DIGITAL REFERENCE//REFERENCE DESK	61	66	92%
HOPE OLSON//BISAC//KNOWLEDGE ORGANIZATION	57	66	86%
ACADEMIC LIBRARY USE//ACADEMIC AND RESEARCH LIBRARIES// ACADEMIC ASSIGNMENTS	48	51	94%

Appendix 1. (*continued*)

SPECIALTY TOPIC	# articles in P_{lis}	# articles in topic	Share of topic in P_{lis}
INTERLENDING//DOCUMENT DELIVERY//ACADEMIC LIBRARIES			
ELECTRONIC BOOKS//E BOOKS//E TEXTBOOK	134	172	78%
OPEN ACCESS//OPEN ACCESS JOURNALS//GOLD OPEN ACCESS	123	218	56%
DOCUMENT DELIVERY//INTERLENDING//INTERLIBRARY LOAN	121	122	99%
ELECTRONIC JOURNALS//ELECTRONIC PERIODICALS//E JOURNALS	79	85	93%
KOHA//INTEGRATED LIBRARY SYSTEMS//WEB SCALE DISCOVERY	72	78	92%
INSTITUTIONAL REPOSITORIES//ACADEMIC AUTHORS// DIGITAL LIBRARY FRAMEWORK	63	69	91%
RESEARCH DATA//DATA SHARING//DATA REPOSITORIES	53	147	36%
INTERFACE CONSISTENCY//ADAPTIVE LIBRARY SERVICES// ALEXANDRIA DIGITAL LIBRARY PROJECT	36	46	78%
CITATION STUDY//COLLECTION ASSESSMENT//ACADEMIC MEDICAL CENTER LIBRARY	31	38	82%
FRBR//DESCRIPTIVE CATALOGUING//FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS FRBR	29	35	83%
RECOMMENDER SYSTEMS//COLLABORATIVE FILTERING// INFORMATION RETRIEVAL			
FOLKSONOMY//SOCIAL TAGGING//COLLABORATIVE TAGGING	57	159	36%
SENTIMENT ANALYSIS//OPINION MINING//SENTIMENT CLASSIFICATION	35	371	9%
RECOMMENDER SYSTEMS//COLLABORATIVE FILTERING// RECOMMENDATION SYSTEM	30	576	5%
RELEVANCE CRITERIA//RELEVANCE JUDGEMENT//TEST COLLECTIONS	25	49	51%
SESSION IDENTIFICATION//QUERY LOG ANALYSIS// QUERY RECOMMENDATION	24	86	28%
COMMUNITY QUESTION ANSWERING//SOCIAL QA// ANSWER RECOMMENDATION	23	54	43%
COLLABORATIVE INFORMATION SEEKING//SEARCH HISTORIES// SOCIAL SEARCH	22	40	55%
EXPERT FINDING//EXPERT SEARCH//ENTITY RETRIEVAL	20	87	23%
MULTI DOCUMENT SUMMARIZATION//TEXT SUMMARIZATION// DOCUMENT SUMMARIZATION	16	146	11%
STEMMING//CROSS LANGUAGE INFORMATION RETRIEVAL// CHARACTER N GRAMS	11	40	28%

Appendix 1. (*continued*)

SPECIALTY TOPIC	# articles in P_{lis}	# articles in topic	Share of topic in P_{lis}
ELECTRONIC HEALTH RECORDS//ELECTRONIC MEDICAL RECORD// MEDICAL INFORMATICS			
NATURAL LANGUAGE PROCESSING//MEDICAL LANGUAGE PROCESSING//CLINICAL TEXT	92	278	33%
HEALTH INFORMATION TECHNOLOGY//ELECTRONIC HEALTH RECORDS//MEANINGFUL USE	46	392	12%
ALERT FATIGUE//CLINICAL DECISION SUPPORT SYSTEMS// CLINICAL DECISION SUPPORT	43	216	20%
CDISC//ISO IEC 11179//CLINICAL RESEARCH INFORMATICS	37	167	22%
PHEWAS//PHENOME WIDE ASSOCIATION STUDY// CLINICAL PHENOTYPE MODELING	34	160	21%
HEALTH INFORMATION EXCHANGE//HEALTH RECORD BANK// REGIONAL HEALTH INFORMATION ORGANIZATIONS	32	154	21%
CPOE//E PRESCRIBING//ELECTRONIC PRESCRIBING	31	218	14%
OPENEHR//LOINC//CLINICAL ARCHETYPES	21	111	19%
SNOMED CT//UMLS//ABSTRACTION NETWORK	13	108	12%
COPY PASTE//CLINICAL DOCUMENTATION//COMPUTER BASED DOCUMENTATION	10	59	17%
ENTERPRISE RESOURCE PLANNING//ENTERPRISE RESOURCE PLANNING ERP// END USER COMPUTING			
IT BUSINESS VALUE//BUSINESS VALUE OF IT//IT INVESTMENT	92	184	50%
STRATEGIC INFORMATION SYSTEMS PLANNING//CHIEF INFORMATION OFFICER//IT GOVERNANCE	70	124	56%
IS RESEARCH//REFERENCE DISCIPLINE//IS DISCIPLINE	65	83	78%
ENTERPRISE RESOURCE PLANNING//ENTERPRISE RESOURCE PLANNING ERP//ERP IMPLEMENTATION	41	196	21%
TOE FRAMEWORK//E COMMERCE ADOPTION//TECHNOLOGY ORGANIZATION ENVIRONMENT FRAMEWORK	38	136	28%
REQUIREMENTS UNCERTAINTY//SYSTEM SUCCESS// SOFTWARE PROJECT RISK	23	81	28%
CAREER ANCHORS//IS PERSONNEL//IT WORKFORCE	23	49	47%
DATA QUALITY//INFORMATION QUALITY MANAGEMENT// INFORMATION QUALITY	15	68	22%
USER SATISFACTION//INFORMATION SYSTEMS SUCCESS//IS SUCCESS	15	81	19%
SUBJECTIVITY STUDY//ACTOR ENGAGEMENT// AGILE ANALYTICS	12	73	16%

Appendix 1. (*continued*)

SPECIALTY TOPIC	# articles in P_{lis}	# articles in topic	Share of topic in P_{lis}
KNOWLEDGE MANAGEMENT//KNOWLEDGE SHARING// OPEN SOURCE SOFTWARE			
KNOWLEDGE SHARING//KNOWLEDGE MANAGEMENT// KNOWLEDGE SHARING BEHAVIOR	151	328	46%
KNOWLEDGE MANAGEMENT//ENTERPRISE BENEFITS// KNOWLEDGE CHAIN	58	101	57%
OPEN SOURCE SOFTWARE//OPEN SOURCE//OPEN SOURCE SOFTWARE OSS	53	230	23%
WIKIPEDIA//COOPERATIVE KNOWLEDGE GENERATION//ENCYCLOPAEDIAS	22	76	29%
PERSONAL INFORMATION MANAGEMENT//ADAPTIVE WINDOW MANAGER//ADVANCED MANAGEMENT OF PERSONAL INFORMATION	22	48	46%
COMMUNITIES OF PRACTICE//ORGANIZING PRACTICES// COMMUNITY OF PRACTICE	19	82	23%
INTELLECTUAL CAPITAL//INTANGIBLE ASSETS//INTELLECTUAL CAPITAL IC	19	96	20%
ENTERPRISE EVOLUTION//KNOWLEDGE CREATION// AUTOMOBILE PROJECT	16	43	37%
EUROPEAN SMES//BARRIERS OF IMPLEMENTATION// CASE STUDY IN SINGAPORE	13	35	37%
BLACK HAT SEO//COMMUNICATIONS ACTIVITIES// CONSUMER COMPARISON	11	28	39%
INNOVATION//PATENTS//OPEN INNOVATION			
PATENT ANALYSIS//PATENT MINING//TECHNOLOGY INTELLIGENCE	41	179	23%
NON PATENT REFERENCES//NON PATENT CITATION//SCIENCE LINKAGE	34	57	60%
PROBABILISTIC ENTROPY//UNIVERSITY INDUSTRY GOVERNMENT RELATIONSHIP//TRIPLE HELIX	32	52	62%
ACADEMIC ENTREPRENEURSHIP//ENTREPRENEURIAL UNIVERSITY// UNIVERSITY SPIN OFFS	27	421	6%
PATENT VALUE//PATENTS//PATENT SYSTEM	27	202	13%
USER INNOVATION//LEAD USERS//INNOVATION CONTESTS	19	226	8%
SOFTWARE ECOSYSTEMS//BUSINESS ECOSYSTEM// MOBILE COMPUTING INDUSTRY	16	81	20%
ABSORPTIVE CAPACITY//POTENTIAL ABSORPTIVE CAPACITY// COMBINATIVE CAPABILITIES	13	102	13%
NATIONAL ELIGIBILITY TEST//RD EFFICIENCY// CHINAS HIGH TECH INNOVATIONS	10	47	21%
INVENTIVE ACTIVITIES//ASSIGNEE//CO PATENT	9	17	53%

Appendix 1. (*continued*)

SPECIALTY TOPIC	# articles in P_{lis}	# articles in topic	Share of topic in P_{lis}
UNIVERSAL SERVICE//TELECOMMUNICATIONS//ACCESS PRICING			
DIGITAL DIVIDE//BROADBAND ADOPTION//BROADBAND	71	150	47%
ACCESS REGULATION//ACCESS PRICING//NEXT GENERATION ACCESS NETWORKS	48	121	40%
TD SCDMA//FORMAL STANDARDS//WAPI	39	63	62%
SPECTRUM AUCTIONS//DIGITAL DIVIDEND//SPECTRUM TRADING	28	68	41%
FIXED MOBILE SUBSTITUTION//MOBILE TELECOMMUNICATIONS// FIXED TO MOBILE SUBSTITUTION	24	56	43%
BILL AND KEEP//TERMINATION RATES//ACCESS PRICING	17	60	28%
UNIVERSAL SERVICE//E RATE//UNIVERSAL SERVICE FUND	17	35	49%
NET NEUTRALITY//NETWORK NEUTRALITY//CONTENT PROVIDERS	17	70	24%
PRICE CAPS//INCENTIVE REGULATION//PRICE CAP REGULATION	12	39	31%
AUSTRALIAN EXPERIENCE//BARRIERS TO TRADE AND INVESTMENT// CAUSAL CHAIN OF REFORM	8	30	27%
HEALTH LITERACY//INTERNET//MHEALTH			
HEALTH LITERACY//NEWEST VITAL SIGN//S TOFHLA	73	544	13%
PERSONAL HEALTH RECORDS//PATIENT PORTAL// PATIENT ACCESS TO RECORDS	39	281	14%
HEALTH INFORMATION SEEKING//HEALTH INFORMATION AVOIDANCE//INFORMATION SEEKING	35	118	30%
ONLINE SUPPORT GROUPS//COMPREHENSIVE HEALTH ENHANCEMENT SUPPORT SYSTEM CHESS// INTERNET CANCER SUPPORT GROUPS	28	220	13%
MHEALTH//MEDICATION REMINDERS//REAL TIME ADHERENCE MONITORING	20	275	7%
INTERNET//HEALTH INFORMATION//ONLINE HEALTH INFORMATION	14	193	7%
TEXT MESSAGING//TEXT MESSAGE//MHEALTH	13	218	6%
DISCERN//INTERNET//QUALITY OF INFORMATION	8	207	4%
INTERNET CHILD HEALTH INFORMATION//ASSESSMENT OF ACUTE DISEASES//AUTISM CEREBRAL PALSY	7	46	15%
READABILITY//PATIENT EDUCATION MATERIALS// FLESCH KINCAID GRADE LEVEL	6	147	4%

Appendix 2. Topics per Specialty – MIS

SPECIALTY TOPIC	# articles in P_{mi}	# articles in topic	Share of topic in P_{mi}
ELECTRONIC HEALTH RECORDS//ELECTRONIC MEDICAL RECORD//MEDICAL INFORMATICS			
NATURAL LANGUAGE PROCESSING//MEDICAL LANGUAGE PROCESSING//CLINICAL TEXT	188	278	68%
HEALTH INFORMATION TECHNOLOGY//ELECTRONIC HEALTH RECORDS//MEANINGFUL USE	127	392	32%
CDISC//ISO IEC 11179//CLINICAL RESEARCH INFORMATICS	112	167	67%
ALERT FATIGUE//CLINICAL DECISION SUPPORT SYSTEMS// CLINICAL DECISION SUPPORT	110	216	51%
NURSING INFORMATION SYSTEM//CLINICAL INFORMATION SYSTEMS//DOCUMENTATION TIME	104	152	68%
OPENEHR//LOINC//CLINICAL ARCHETYPES	89	111	80%
HEALTH INFORMATION EXCHANGE//HEALTH RECORD BANK// REGIONAL HEALTH INFORMATION ORGANIZATIONS	87	154	56%
CPOE//E PRESCRIBING//ELECTRONIC PRESCRIBING	84	218	39%
SNOMED CT//UMLS//ABSTRACTION NETWORK	76	108	70%
PHEWAS//PHENOME WIDE ASSOCIATION STUDY// CLINICAL PHENOTYPE MODELING	49	160	31%
HEALTH LITERACY//INTERNET//MHEALTH			
PERSONAL HEALTH RECORDS//PATIENT PORTAL//PATIENT ACCESS TO RECORDS	112	281	40%
INTERNET//HEALTH INFORMATION//ONLINE HEALTH INFORMATION	48	193	25%
ONLINE SUPPORT GROUPS//COMPREHENSIVE HEALTH ENHANCEMENT SUPPORT SYSTEM CHESS//INTERNET CANCER SUPPORT GROUPS	37	220	17%
MEDICAL APP//APPS//SMARTPHONE	33	165	20%
MHEALTH//MEDICATION REMINDERS//REAL TIME ADHERENCE MONITORING	32	275	12%
TWITTER MESSAGING//INFOVEILLANCE//INFODEMIOLOGY	31	113	27%
E PROFESSIONALISM//SOCIAL MEDIA//TWITTER MESSAGING	30	274	11%
TEXT MESSAGING//TEXT MESSAGE//MHEALTH	29	218	13%
MOBILE APPS//APPS//MHEALTH	25	126	20%
PHYSICIAN RATING WEBSITE//RATING SITES//QUALITY TRANSPARENCY	23	61	38%

Appendix 2. (*continued*)

SPECIALTY TOPIC	# articles in P_{mi}	# articles in topic	Share of topic in P_{mi}
HEALTH TECHNOLOGY ASSESSMENT//EQ 5D//PRIORITY SETTING			
HEALTH TECHNOLOGY ASSESSMENT//HOSPITAL BASED HTA//MINI HTA	58	118	49%
EQ 5D//SF 6D//EQ 5D 5L	38	411	9%
VALUE OF INFORMATION//OPTIMAL TRIAL DESIGN// VALUE OF INFORMATION ANALYSIS	29	95	31%
HEALTH TECHNOLOGY ASSESSMENT//INSTITUTE FOR QUALITY AND EFFICIENCY IN HEALTH CARE//FOURTH HURDLE	28	153	18%
DYNAMIC TRANSMISSION//HALF CYCLE CORRECTION// COST EFFECTIVENESS MODELING	26	82	32%
STRENGTH OF PREFERENCES//IN PERSON INTERVIEW// MULTI CRITERIA DECISION ANALYSIS	15	67	22%
COST EFFECTIVENESS RATIOS//NET HEALTH BENEFIT// COST EFFECTIVENESS ACCEPTABILITY CURVES	14	74	19%
COVERAGE WITH EVIDENCE DEVELOPMENT// MEDICARE COVERAGE//RISK SHARING AGREEMENTS	14	91	15%
HORIZON SCANNING SYSTEMS//HORIZON SCANNING// EARLY AWARENESS AND ALERT SYSTEMS	12	20	60%
COMPARATIVE EFFECTIVENESS RESEARCH//PATIENT CENTERED OUTCOMES RESEARCH//ELECTRONIC CLINICAL DATA	10	158	6%
ADAPTIVE DESIGN//INTERIM ANALYSIS//DOSE FINDING			
ADAPTIVE DESIGN//GROUP SEQUENTIAL TEST// GROUP SEQUENTIAL DESIGN	58	221	26%
CONTINUAL REASSESSMENT METHOD//DOSE FINDING// DOSE FINDING STUDIES	50	200	25%
TWO STAGE DESIGN//PHASE II DESIGN//PHASE II CLINICAL TRIALS	28	113	25%
FAMILYWISE ERROR RATE//GATEKEEPING PROCEDURE//MULTIPLE TESTS	27	119	23%
SCORE INTERVAL//BINOMIAL PROPORTION//BINOMIAL DISTRIBUTION	18	140	13%
NONINFERIORITY MARGIN//NON INFERIORITY//NON INFERIORITY TRIAL	16	112	14%
MONOTONE MISSING//DISCRETE TIME LONGITUDINAL DATA// INDEPENDENT MISSING	12	37	32%
MINIMUM EFFECTIVE DOSE//MCP MOD//WILLIAMS TEST	12	57	21%
META ANALYTIC PREDICTIVE//EPSILON INFORMATION PRIOR// COMPUTATIONALLY INTENSIVE METHODS	9	42	21%
MULTIREGIONAL CLINICAL TRIAL//BRIDGING STUDY// MULTIREGIONAL TRIAL	8	68	12%

Appendix 2. (*continued*)

SPECIALTY TOPIC	# articles in P_{mi}	# articles in topic	Share of topic in P_{mi}
MISSING DATA//MULTIPLE IMPUTATION//GENERALIZED ESTIMATING EQUATIONS			
GENERALIZED ESTIMATING EQUATIONS//QUASI LEAST SQUARES//GEE	23	104	22%
MULTIPLE IMPUTATION//MISSING DATA//PREDICTIVE MEAN MATCHING	23	218	11%
JOINT MODEL//SHARED PARAMETER MODEL//DYNAMIC PREDICTIONS	22	112	20%
CONCORDANCE CORRELATION COEFFICIENT//TOTAL DEVIATION INDEX//COEFFICIENT OF INDIVIDUAL AGREEMENT	19	66	29%
PATTERN MIXTURE MODEL//MISSING NOT AT RANDOM//MISSING DATA	19	137	14%
ZERO INFLATION//ZERO INFLATED MODELS//OVERDISPERSION	16	142	11%
CENSORED COVARIATE//CENSORED PREDICTOR// TWO PART STATISTICS	13	41	32%
REGRESSION CALIBRATION//MEASUREMENT ERROR//CORRECTED SCORE	12	105	11%
INFORMATIVE CLUSTER SIZE//WITHIN CLUSTER RESAMPLING// CLUSTERED OBSERVATIONS	10	43	23%
DOUBLE ROBUSTNESS//AUGMENTED INVERSE PROBABILITY WEIGHTING AIPW//MISSING AT RANDOM	10	67	15%
COMPETING RISKS//INTERVAL CENSORING//COUNTING PROCESS			
INTEGRATED DISCRIMINATION IMPROVEMENT//NET RECLASSIFICATION IMPROVEMENT//DECISION ANALYTIC MEASURES	28	110	25%
MULTISTATE MODEL//ILLNESS DEATH PROCESS//AALEN JOHANSEN ESTIMATOR	23	111	21%
COMPETING RISKS//CUMULATIVE INCIDENCE FUNCTION// CAUSE SPECIFIC HAZARD	20	117	17%
RECURRENT EVENTS//PANEL COUNT DATA// INFORMATIVE OBSERVATION TIMES	19	174	11%
EXPLAINED VARIATION//TIME DEPENDENT ROC//C INDEX	19	68	28%
CURE RATE MODEL//CURE MODEL//LONG TERM SURVIVAL MODELS	17	95	18%
SURROGATE ENDPOINT//PRENTICE CRITERION// LIKELIHOOD REDUCTION FACTOR	13	84	15%
INTERVAL CENSORING//CURRENT STATUS DATA// INTERVAL CENSORED DATA	13	127	10%
CASE COHORT DESIGN//CASE COHORT//CASE COHORT STUDY	12	77	16%
FRAILTY MODEL//CORRELATED FAILURE TIMES// CROSS RATIO FUNCTION	12	104	12%

Appendix 2. (*continued*)

SPECIALTY TOPIC	# articles in P_{mi}	# articles in topic	Share of topic in P_{mi}
EVIDENCE BASED MEDICINE//PUBLICATION BIAS//ABSTRACT			
MULTIVARIATE META ANALYSIS//DERSIMONIAN LAIRD ESTIMATOR//MANDEL PAULE ALGORITHM	40	150	27%
MEDICAL EPISTEMOLOGY//EVIDENCE BASED MEDICINE// EVIDENCE IN MEDICINE	32	77	42%
MIXED TREATMENT COMPARISON//NETWORK META-ANALYSIS// MULTIPLE TREATMENTS META ANALYSIS	26	198	13%
MEDLINE//EMBASE//LITERATURE SEARCHING	11	112	10%
NUMBER NEEDED TO TREAT//ABSOLUTE RISK REDUCTION// NUMBER NEEDED TO TREAT NNT	9	52	17%
TRIAL REGISTRATION//CLINICALTRIALSGOV//PUBLICATION BIAS	7	250	3%
PUBLICATION BIAS//FUNNEL PLOT//SMALL STUDY EFFECTS	6	64	9%
JOURNAL CLUB//EVIDENCE BASED MEDICINE EDUCATION//FRESNO TEST	6	155	4%
AWARENESS SCORE//CHIROPRACTIC QUESTIONNAIRES// COMMUNITY OF PRACTICE KNOWLEDGE NETWORKS	6	45	13%
CONFLICT OF INTEREST//EDITORIAL ETHICS//CONFLICTS OF INTEREST	6	189	3%
PATIENT SAFETY//MEDICATION ERRORS//MEDICAL ERRORS			
MEDICATION ERRORS//SMART PUMPS//MEDICATION ADMINISTRATION ERRORS	25	281	9%
BAR CODE MEDICATION ADMINISTRATION//BAR CODED MEDICATION ADMINISTRATION//SCANNING COMPLIANCE	24	82	29%
SIGN OUT//HANDOFF//HANDOVER	19	334	6%
VOCERA//HOSPITAL COMMUNICATION SYSTEMS//PAGERS	19	66	29%
INTERRUPTION//DISTRACTIONS//TASK SEVERITY	13	162	8%
MEDWISE//THREAT AND ERROR MANAGEMENT TEM// USER CONFIGURABLE EHR	11	32	34%
INCIDENT REPORTING//MEDICATION INCIDENTS//ERROR REPORTING	8	155	5%
MEDICAL DEVICE DESIGN//INSTITUTIONAL DECISION MAKING//USER COMPUTER	8	34	24%
NON TECHNICAL SKILLS//TEAMWORK//TEAM TRAINING	7	317	2%
TRIGGER TOOL//GLOBAL TRIGGER TOOL//PREVENTABLE HARM	6	182	3%

Appendix 2. (*continued*)

SPECIALTY TOPIC	# articles in P_{mi}	# articles in topic	Share of topic in P_{mi}
TELEMEDICINE//TELEHEALTH//TELEPATHOLOGY			
TELEHEALTH//TELECARE//TELEHEALTHCARE	32	188	17%
TELEMONITORING//HOME TELEMONITORING//TETEMONITORING	27	155	17%
ELDERCARE TECHNOLOGY//HOME BASED CLINICAL ASSESSMENT//PASSIVE INFRARED PIR MOTION DETECTORS	19	90	21%
TELE ECHOGRAPHY//TELESONOGRAPHY//TELE ULTRASOUND	10	48	21%
MOBILE TELEMEDICINE//MOBILE CARE//TIME FREQUENCY ENERGY DISTRIBUTIONS	7	34	21%
CHRONIC DISEASE METHODS THERAPY//CRITICAL PATHWAYS MESH//HEATH CARE PRACTICES	7	24	29%
TELEREHABILITATION//TELEPRACTICE//REMOTE ASSESSMENT	6	90	7%
TELE EEG//INITIATE BUILD OPERATE TRANSFER STRATEGY// INTERNATIONAL VIRTUAL E HOSPITAL FOUNDATION	5	69	7%
ISI WEB OF SCIENCE DATABASE//LISTENING STYLES// NON ADHERENCE FACTORS	5	24	21%
TELEDERMATOLOGY//MOBILE TELEDERMATOLOGY// STORE AND FORWARD	4	135	3%
CAUSAL INFERENCE//PROPENSITY SCORE//PRINCIPAL STRATIFICATION			
PROPENSITY SCORE//OBSERVATIONAL STUDY//COVARIATE BALANCE	34	229	15%
PRINCIPAL STRATIFICATION//NONCOMPLIANCE//CAUSAL INFERENCE	32	159	20%
MARGINAL STRUCTURAL MODELS//TARGETED MAXIMUM LIKELIHOOD ESTIMATION//CAUSAL INFERENCE	29	221	13%
DYNAMIC TREATMENT REGIMES//ADAPTIVE TREATMENT STRATEGIES//OPTIMAL TREATMENT REGIME	19	127	15%
MENDELIAN RANDOMIZATION//MENDELIAN RANDOMISATION// ALLELE SCORES	11	121	9%
PROPENSITY SCORE CALIBRATION//PROBABILISTIC BIAS ANALYSIS//NONDIFFERENTIAL	4	50	8%
RANDOMIZATION INFERENCE//MATCHED SAMPLING//FINE BALANCE	4	56	7%
INSTRUMENTAL VARIABLES//PHYSICIAN PRESCRIBING PREFERENCE//PHYSICIANS PRESCRIBING PREFERENCE	4	64	6%
MARGINAL TREATMENT EFFECT//CORRELATED RANDOM COEFFICIENT MODEL//LOCAL INSTRUMENTAL VARIABLES	1	51	2%
UNCONFOUNDEDNESS//PROPENSITY SCORE MATCHING// SELECTION ON OBSERVABLES	1	204	0%