03/2013

UNCOVER is an FP7-funded project under Contract No 282574

AIT
DUK UNC

Edgar Schiebel

Brigitte Palensky

Maria-Elisabeth Züger

Data sources for bibliometric analysis

UNCOVER project deliverable D2.1

This report should be cited as follows:
Schiebel, E. Palensky, B., Züger, M.-E., Deliverable D2.1 of the UNCOVER FP7-funded project under contract number 282574: Data sources for bibliometric analysis

This deliverable elaborates on the objectives “Framing of hypotheses, definition of suitable and measurable bibliometric features, and definition of statistical indicators (based on features)” and “Extraction, pre-processing, and standardization of data from information sources”. It presents hypotheses for how to overcome or at least shed light into some aspects of publication bias that can be elaborated with bibliometric approaches. For this purpose different data bases and datasets with different field content are prepared and condensed. Available data sources for the calculation of bibliometric indicators and science maps are presented.

This deliverable was prepared for the UNCOVER project consortium:

AIT Austrian Institute of Technology
Department for Foresight & Policy Development
Coordinating partner

Vienna, Austria

DUK Danube University Krems
Department for Evidence-based Medicine and Clinical Epidemiology

Krems, Austria

UNC University of North Carolina at Chapel Hill
Gillings School of Global Public Health

Chapel Hill, North Carolina, United States

Contact
Dr. Manuela KIENEGGER
Technology Management
Department for Foresight & Policy Development
AIT Austrian Institute of Technology GmbH
Donau-City-Straße 1
A-1220 Vienna
Austria

T +43(0) 50550-4530
F +43(0) 50550-4599

manuela.kienegger@ait.ac.at

www.ait.ac.at

Data sources for bibliometric analysis

.

This deliverable was prepared by:

Edgar Schiebel
Brigitte Palensky
Maria-Elisabeth Züger

Other contributions to the deliverable by:

Dirk Holste
Manuela Kienegger
Silvia Steinbrunner
Michaela Strobelberger

Content

Table Index

Figure Index

1 Executive Summary

The objectives "Framing of hypotheses, definition of suitable and measurable bibliometric features, and definition of statistical indicators (based on features)" and "Extraction, pre-processing, and standardization of data from information sources" are pursued in this deliverable. It presents hypotheses for how to overcome or at least shed light into some aspects of publication bias that can be elaborated with bibliometric approaches. For this purpose different data bases and datasets with different field content are prepared and condensed. Available data sources for the calculation of bibliometric indicators and science maps are presented.

Based on a framework of registered or non-registered studies, published or not published trial results we formulated the following hypotheses:

- The registration ID for clinical trials is cited in publications.
- The number of publications with a registration ID is growing.
- If editors insist on publications referring to registered studies, the number of such publications grows.
- Incentives make it more likely for authors and editors to publish registered studies than not registered studies.

Following the Hypotheses base on bibliometric approaches we formulated:

- The citation rate of publications with a registration code deriving from study registers is higher than for publications on clinical trials without a registration number. Higher citation rates are considered an incentive for editors and publishing trialists.
- Journals with a publication policy to only accept publications with a registration number have an influence on the growing number of publications related to registered studies.
- Science maps broaden the view on research issues in systematic reviews and meta-analyses and allow for a more comprehensive selection of relevant literature.

For the verification or falsification of the hypotheses we use relational bibliometric techniques: co-authorships, co-citation analysis and bibliographic coupling and basic indicators such as the number of publications per year and citation rates of single publications.

Three databases, PubMed, Web of Science and clinical trials.gov, were combined for the analysis of three cases (medical topics) that had previously been used for a Systematic Review by the Danube University Krems (DUK):

- Set-I, Antidepressants for Adult Repression (MMA) and
- Set-II, Targeted Immune Modulators (TIM), or

- Set-II, Deceases of the Cardiovascular System (CCVS)

The search in the different data bases for the three medical topics is summarized in the following table:

No.	Database	MMA	TIM	CCVS
(1)	PubMed search	785 (Publ.)	522 (Publ.)	
(2)	WoS overlap with 1 by publication DOI	742 (Publ.)	511 (Publ.)	
(3)	WoS search by topic	3,227 (Publ.)	10,713 (Publ.)	
(4)	Overlap 3 with 2	357 (Publ.)	387 (Publ.)	
(5)	ClinicalTrials.gov	651 (studies)	547 (studies)	
(6)	WoS search			2,727 (Publ.)

The search strategy (1) used in the previously published systematic reviews of DUK in the PubMed database delivered for MMA and for TIM 785 and 522 publications, respectively, in PubMed. A more general search in WoS by topic resulted in 3,220 and 10,713 publications, respectively. The overlap of the WoS search by topic with the PubMed search shows only 357 and 547 publications, respectively.

The information of the different databases was combined, condensed and prepared for the analysis of the data that is part of the UNCOVER Deliverable 2.2.

2 Introduction

2.1 Background

The UNCOVER project is a direct contribution to overcome non-publication of clinical studies that have been designed and executed as randomized controlled trials (RCTs).

UNCOVER's aim is three-fold:

- to apply established and develop novel, solid, and useful methods for fact-finding and interventions into the socio-economic system defined by causes and sources of the publication bias;
- to engage with stakeholders and identify strategies, barriers, and facilitating factors associated with the publication bias and its consequences; and
- to synthesize lessons learned and recommend feasible measures to deal with the publication bias.

RCTs are currently best practice to avoid or minimize both systematic and random errors in clinical studies. They provide the best utility as input to systematic medicinal reviews, one cornerstone of evidence-based medicine (EbM) for improved safety and efficacy / effectiveness of patient outcomes, and their end-users.

That is guaranteed when RCTs are both correctly registered and published at least once. Because non-publication, as well as publication with time delay of RCTs, may decisively reduce the advantage of such systematic reviews of drugs, medical devices or procedures, it affects the knowledge base. Therefore, in a perspective way, this project contributes pro better allocation of funds to sponsor studies and patient value, and contra duplication of work and patients risk.

The issues of the publication bias are treated with quantitative, qualitative and participatory means in an interdisciplinary approach in areas with little or no lines of evidence as to how they perform in practice:

1. Framing the publication bias in terms of EbM and system's theory (including stakeholder mapping) to both acknowledge and reduce the complexity of the problem and focus on the main players in publishing studies as well as their strategies.
2. Objective, systematic and balanced identification of key opinion leaders, as well as measures (law, regulations, policies, practices, guidelines, methods, and tools) to overcome bias, from documents and sites by bibliometric means and comprehensive site searches on the world-wide web.

3. Systematic review of current measures substantiated by own experience ("inside-out") as well as inclusion of experts and external knowledge of international methods groups ("outside-in") in the field of systematic reviews and meta-analyses.
4. Design of interviews (telephone, or face-to-face) with editors and other stakeholders based on stakeholder mapping/analysis to reflect measures in terms of experiences, own strategies and existing conflict of interests.
5. Development of needed software solutions for the demonstration and treatment of unpublished studies on statistical meta-analyses.
6. Recommendations for the implementation of feasible measures and milestones, as well as open gaps addressed by new research, to overcome non-publication.

UNCOVER will thus both provide viable solutions for encountering publication bias.

2.2 Objectives of WP2

The objectives of WP2 are:

- Framing of hypotheses, definition of suitable and measurable bibliometric features, and definition of statistical indicators (based on features)
- Extraction, pre-processing, and standardization of data from information sources
- Investigation of the influence of registered vs. non-registered studies on the obtained bibliometric profile in the case of both a systematic review and a comprehensive thematic compilation of medicinal research studies
- Interpretation of characteristic features and conclusions given the current measures against publication bias

Work package 2 is structured in three tasks:

- Task 2.1 Definition of features and indicators for bibliometric analysis and data extraction
- Task 2.2 Bibliometric analysis of characteristic features distinctive between registered v. non-registered studies and
- Task 2.3 Characteristic bibliometric features of publication bias and conclusions

2.3 Aim of the Task 2.1

This deliverable is related to the two objectives: “Framing of hypotheses, definition of suitable and measurable bibliometric features, and definition of statistical indicators (based on features)” and “Extraction, pre-processing, and standardization of data from information sources”. It aims to establish hypotheses for how to overcome or at least shed light into some aspects of publication bias that can be elaborated with bibliometric approaches. For this purpose different data bases (PubMed, Web of Science and clinical trials.gov) and datasets with different field content have to be prepared and condensed. Available data sources for the calculation of bibliometric indicators and science maps will be presented.

For the verification or falsification of the hypotheses we will use relational bibliometric techniques such as co-authorships, co-citation analysis and bibliographic coupling and basic indicators such as the number of publications per year and citation rates of single publications. In addition, the tree databases will be combined for the analysis of three cases (medical topics) that had previously been used for a Systematic Re-view by the Danube University Krems (DUK).

While Deliverable 2.1 describes the preparatory work for all the analyses, the results of the analyses will be part of the UNCOVER Deliverable 2.2.

2.4 Organization of This Report

This report is organized as follows. Section 3 deals with the formulation of hypotheses concerning publications on randomized controlled trials that have either been registered or not, and has also a focus on whether measures to register trials have had an effect on publication practice in the last years. A second class of hypotheses is related to the literature basis that is used for systematic reviews and meta-analyses. The methodological framework about bibliometric approaches is described in Section 4. The data sources, search strategies and data structures are discussed in Section 5. Results and discussion are presented in Section 6 and 7, respectively.

3 Framing of Hypotheses

Bibliometrics deals with the analysis of literature. In the context of publication bias two aspects are of interest: (1) publication of results derived from registered or non-registered studies, and (2) finding the relevant literature for systematic reviews and meta-analyses as well as getting a broader view on research issues. In the following we formulate hypotheses and central questions for our further analyses.

3.1 Hypotheses for publications deriving from registered and non-registered studies

The hypotheses are focused on the interrelation of publications based on randomized clinical studies that have been either registered or not. The underlying framework for building these hypotheses is given in Figure 3.1.

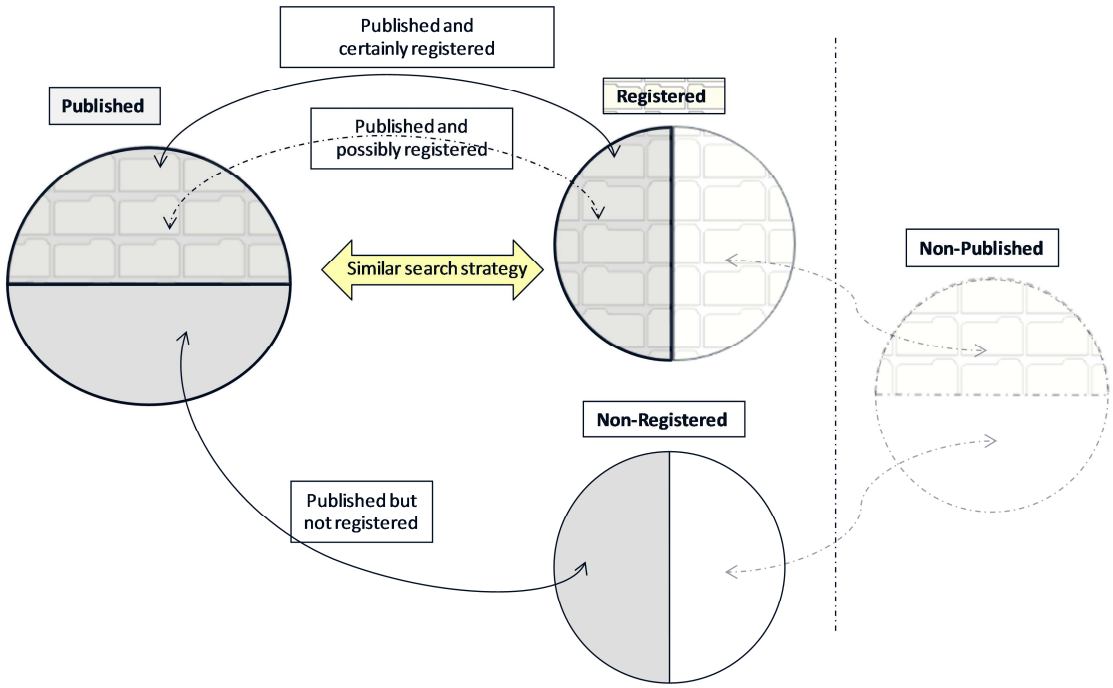

Figure 3.1: Relation between published, (non-published), registered, and non-registered trials. The circles to the left of the vertical dotted line represent a set of publications on the one hand, or registered or non-registered clinical trials on the other hand. They are all to a specific topic *("Similar search strategy"*). Publication is indicated by a *grey filling*, registration is indicated by a *repeating pattern of folders*. The information about whether a publication is based on a registered or on a non-registered trial may be transparent ("Published and certainly registered") or hidden ("Published and possibly registered").

The relationships are as follows:

- clinical trials are either registered or not;
- registered trials are either published or not;
- publications are derived from clinical trials that have been either registered or not;
- the results of unregistered trials can remain unpublished, etc.

Information on the relation between publications and registered clinical trials can be direct and easily accessible, or, more or less, hidden and therefore difficult or even impossible to reveal. For example, direct relation between a publication and a registered trial can be easily identified in a publication through reference to an unambiguous and unique identification number of the trial, such as those issued by trial registries, e.g. the ISRCTN (Interna-

tional Standard Randomised Controlled Trial Number) for international standardized RCTs, or the NCT for clinical trials registered with ClinicalTrials.gov.

For any publication, the registration number can be placed at various positions either in the literature database that stores information on the publication or the publication itself. If a literature database contains a database field specifically dedicated to a registration number, the relationship between a publication and the registration is easily accessible in terms of bibliometric analysis. If the registration number is integrated at a specific position in the text of the publication, for instance, cited always at the end of the abstract, then the relationship between a publication and the registration of the related study is more difficult to access. If the number cannot always be located at a specific position in the publication, but can be expected to be found anywhere in the text of the publication, the accessibility of the relation becomes even more difficult. In case the registration number it not provided at all, but the title of the clinical trial instead, the relation between a publication and the study can hardly be established by bibliometrically means.

We formulate the following hypotheses:

- The registration ID for clinical trials is cited in publications.
- The number of publications with a registration ID is growing.
- If editors insist on publications referring to registered studies, the number of such publications grows.
- Incentives make it more likely for authors and editors to publish registered studies than not registered studies.

3.2 Hypotheses based on Bibliometrics

In the last years relational bibliometric research has made progress in analysing literature, especially concerning the delineation and identification of research issues in broader fields by co-citation analysis and bibliographic coupling as well as the visualization of large datasets. Network analyses based on co-authorships and co-affiliations are used to examine communities in science. These approaches will be also used here to identify communities of researchers who publish results derived from studies that have been either registered or not to delineate similar publications within a research topic of a systematic review and broaden the view on publications on related research.

Citation analysis is applied to examine the citation rates of publications within the context of registered or non-registered studies. As impact factors are highly relevant for editors and authors it will be examined if publications from registered studies are more highly cited than others.

Bibliometric approaches are used for the first time in the context of publication bias and systematic reviews. The core idea is to apply the bibliometric approach for selected research questions on publication bias. We set up the following hypotheses:

- The citation rate of publications with a registration code deriving from study registers is higher than for publications on clinical trials without a registration number. Higher citation rates are considered an incentive for editors and publishing trialists.
- Journals with a publication policy to only accept publications with a registration number have an influence on the growing number of publications related to registered studies.
- Science maps broaden the view on research issues in systematic reviews and meta-analyses and allow for a more comprehensive selection of relevant literature.

4 Methods and bibliometric features

All analyses that involve relational data structures will be computed and analysed by using the software BIBTECHMON™ a bibliometric monitoring system to generate, illustrate and study the interrelations of co-authorships, co-citations or content similarity. BIBTECHMON™ is based on relational mapping, which gives rise to networks (a set of nodes with edges between nodes) of authors, institutions, and other objects. The software allows to simultaneously capture all significant occurring relations, their positioning as well as the overall structure including their development over time. Additional features are: extraction of keywords, density maps of similar objects, graphical retrieval, etc.

Co-occurrence analysis and its visualization in networks and science maps is the approach to illustrate and study co-authorships, co-citations and content similarity. Figure 4.1 shows the basic idea: co-publishing authors occur one more times together in publications. For example, if two authors have 7 common publications they share a co-occurrence of 7. A co-occurrence matrix of all authors of a data set of publications is transformed to a similarity matrix by the Jaccard Index. The visualization of the network is done by a spring model. Networks are drawn with nodes and edges. The size of nodes is proportional to the number of the author's publications; the thickness of the edge is proportional to the Jaccard Index. Networks are analysed by indicators of the social network analysis (SNA).

Maps of similar publications use common references for the calculation of co-occurrences and similarity. This approach is called bibliographic coupling. The spring model places similar publications close to each other in a two-dimensional space. A local agglomeration of similar publications accumulates all publications with very similar reference lists. Publications with different research issues of a given broader research field are grouped together

and are separated from other accumulations with a different kind of “knowledge base”. Such accumulations of new publications are also referred to as “research fronts”. Different areas in the landscape are visualized by a local density function. This density function is calculated by the local number of publications weighted by the strength of the edges. This approach allows the delineation of different research issues or approaches in a common field like “Antidepressants for adult Repression”. Bibliographic coupling also allows for fast retrieval of similar publications.

ID	Authors	Organisations	References	Pub. Year
WOS:000295112000014	Bobo, WV; Chen, H; Trivedi, MH; Stewart, JW; ...	Vanderbilt Univ, Sch Med, Dept Psychiat, Nashville, TN 37212 USA; KK Womens & Childrens Hosp, Mental Wellness Serv, Singapore, Singapore; …	Rush AJ, 2003, BIOL PSYCHIAT, V54, P573, DOI 10.1016/S0006-3223(03)01866-8; Blier P, 2010, AM J PSYCHIAT, V167, P281, DOI 10.1176/appi.ajp.2009.09020186; …	2011
WOS:000225508200016	Fava, M; Alpert, JE; Carmin, CN; Wisniewski, SR; ...	Massachusetts Gen Hosp, Depress Clin & Res Program, Boston, MA 02114 USA; Univ Illinois, Chicago, IL USA; …	Rush AJ, 2003, BIOL PSYCHIAT, V54, P573, DOI 10.1016/S0006-3223(03)01866-8; Fava M, 2003, PSYCHIAT CLIN N AM, V26, P457, DOI 10.1016/S0193-953X(02)00107-7; …	2004

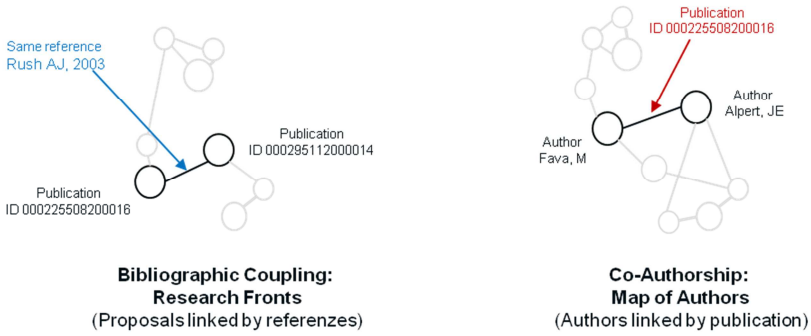

Figure 4.1: Principles of bibliographic coupling and co-authorship.

Citation rates for single publications are provided by the Web of Science (WoS) database. Here, we are to combine three different databases: PubMed, clinical trials.gov and Web of Science to calculate citation rates of publications from registered studies and of those from unregistered studies. The trivial statistic calculations are done by queries with MS-Access and calculations with MS-Excel.

Bibliometric features and indicators to examine the mentioned hypotheses are:

- Scientific activity measured by the number of publications
- Citation analysis
- Co-author analysis
- Bibliographic coupling
- Co-citation analysis
- Combination and condensing different data sources

5 Data sources

Three datasets relating to medical topics that had been previously used for Systematic Reviews by the Danube University Krems will be analysed bibliometrically. The three analysed cases are: “Antidepressants for adult Repression”, “Targeted Immune Modulators” and "Diseases of the cardio vascular system”.

In this task the relevant literature for the mentioned three cases is collected, databases described and differences in coverage and providing the kind of information is part of this task. The analysis and interpretation of the science maps is part of Task 2.2 (analyses) and 2.3 (lessons learned) and is documented in Deliverable D2.2

The objectives are pursued by implementing the following steps:

- Description of the databases
- Search strategy
- Comparison of different kind of information in the databases
- Coverage of information about the registration of studies
- Compilation of publication data sets from three published systematic reviews:
    - Set-I, Antidepressants for Adult Repression (2007) and
    - Set-II, Targeted Immune Modulators (2010), or alternatively
    - Set-III, Deceases of the Cardiovascular System
- Compilation of additional data sets from clinical trial registries (including but not limited to Clinicaltrials.gov) and other data sources (ISI Web of Science)

5.1 Databases

In evidence-based medicine, meta-analyses and systematic reviews rely predominantly on databases such as PubMed and ClinicalTrials.gov and sometimes grey literature. Publications of clinical trials (CT) are indexed in the database PubMed, which has a medical focus. The features of this database are guided by the aim to provide information relevant for medical research. Clinical trials are registered in databases such as clinicaltrials.gov, which provide specific information on facts concerning the trial procedure such as name of the CT, sponsors, dates, protocols, location and other CT specific details.

In contrast, bibliometric approaches usually rely on databases such as Web of Science or Scopus. These databases focus on features that provide bibliographic information such as title, authors, affiliations, standardized references, counts of citations, etc.

The current study uses two data sets on selected medical topics as test cases. These data sets originate from previously carried out Systematic Reviews that were provided by the partner DUK.

The data sets for the two topics are, respectively:

(1) Antidepressants for Adult Repression (MMA), based on the systematic review "*Effectiveness of second generation antidepressants in pharmacological treatment of adult depression*" published in the year 2007, and

(2) Targeted Immune Modulators (TIM), based on the systematic review "*Drug class review on targeted immune modulators*" published in the year 2010.

We had foreseen a third analysis about "Deceases of the Cardiovascular System" as a possible alternative for TIM, however, it was not carried out.

For each topic, the data was obtained from the PubMed database, the ClinicalTrials.gov registry and the ISI Web of Science by using a similar search strategy (as far as possible given the different nature of the databases).

PubMed is a widely used meta-database for publications on medical topics. It was developed by the National Center for Biotechnology Information (NCBI) which describes PubMed as follows: "*A service of the National Library of Medicine that provides access to over 17 million citations from MEDLINE and additional life sciences journals. PubMed includes links to many sites providing full text articles and other related resources.*"[1]

ClinicalTrials.gov is a registry for clinical trials. "*ClinicalTrials.gov is a Web-based resource that provides […] easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions. The Web site is maintained by the National Library of Medicine (NLM) at the National Institutes of Health (NIH). […] Studies are generally submitted to the Web site (that is, registered) when they begin, and the information on the site is updated throughout the study. In some cases, results of the study are submitted after the study ends. This Web site and database of clinical studies is commonly referred to as a 'registry' and 'results database.'*"[2]

Web of Science is a Web of Knowledge database, which is provided by Thomson Reuters. While the Web of Knowledge "*is a research platform that gives you access to objective content and powerful tools to search, track, measure and collaborate in the sciences, social sciences, arts, and humanities*", the Web of Science is an academic citation indexing and

1 http://www.ncbi.nlm.nih.gov/About/tools/restable_lit.html
2 http://www.clinicaltrials.gov/ct2/about-site/background

search service offered by the Web of Knowledge and provides structured bibliographic information. [3]

PubMed focuses mainly on medicine and biomedical sciences, whereas Web of Science covers most scientific fields and could give a broader view also on medical issues from an interdisciplinary point of view. Additionally, it provides citation analysis and standardized references for bibliographic coupling.

Each of the databases has strengths and weaknesses when judging the quality and usefulness of the search results for bibliometric analysis. For instance, PubMed does not deliver some pieces of information (e.g. the full reference list of each publication, or a list of all the authors' affiliations of a publication (instead of just the affiliation of the first author)) which ISI Web of Science does. Vice-versa, PubMed delivers some data that ISI does not, for instance, a field where associated clinical trials are referred to by explicitly giving the registry number. However, the field does not contain many entries (for instance, only about 8 % of the total number of publications for the MMA topic have an entry for the registry number)[4].

The weaknesses of the PubMed results regarding their usefulness for bibliometric analysis were the main reason why an additional search in ISI Web of Science was carried out in the first place. Of course, it would be desirable if there were a sufficiently big overlap of publications between the two searches (that could be determined with reasonable effort) so that part of the analyses could be carried out with ISI web of science data instead of PubMed data.

5.2 Search Strategy

The search strategies are based on the original search strategies carried out for the systematic reviews mentioned above. For each topic, Antidepressants (MMA) and Target Immune Modulators (TIM), we conducted a search in 3 databases: PubMed, ClinicalTrials.gov and Web of Science. A description on the data sources is available in Section 5.1. The following two sections (5.2.1 and 5.2.2) deal with search terms, search settings and the number of documents retrieved.

3 http://wokinfo.com/about/whatitis/;
http://wokinfo.com/products_tools/multidisciplinary/webofscience/

4 This immediately leads to the question of the meaning of this fact, that is, whether the number of publications linked to a clinical trial actually is higher (or not) and how to find out.

5.2.1 Data Set-I: Antidepressants for Adult Repression (MMA)

Source 1: PubMed

The search was conducted on 5 December 2011. The search strategy is made up of 4 blocks including search terms on drugs and agents, disorders, adverse effects as well as methods and trials, which are finally combined by conjunction. Table 5.1 shows the full search strategy and used search terms. In total, 785 publications were acquired.

Table 5.1: Search Strategy of Antidepressants (MMA), Source: PubMed. In addition to the search terms, the used search field is stated in square brackets after each term. Any restrictions during the search are mentioned accordingly with the label 'limits'. The column 'Hits' lists the number of documents retrieved by each search step.

Search Step	Search Strings and Boolean Operations	Hits
1	"Antidepressive Agents, Second-Generation"[MeSH] OR "Fluoxetine"[MeSH] OR "Sertraline"[MeSH] OR "Paroxetine"[MeSH] OR "Citalopram"[MeSH] OR "Fluvoxamine"[MeSH] OR "Bupropion"[MeSH] OR "nefazodone"[Substance Name] OR "mirtazapine"[Substance Name] OR "venlafaxine"[Substance Name] OR "escitalopram"[tw] OR "duloxetine"[Substance Name] OR "Trazodone"[MeSH] OR "O-desmethylvenlafaxine"[Substance Name] OR desvenlafaxine	21,532
2	"Depressive Disorder"[MeSH] OR "Depressive Disorder, Major"[MeSH] OR "Dysthymic Disorder"[MeSH] OR ("depression"[tiab] AND "involutional"[tiab]) OR ("subsyndromal"[tiab] AND "depressive disorder"[tiab])	69,948
3	#1 AND #2 *Limits: Humans, English, All Adult: 19+ years*	4,218
4	"Quality of Life"[MeSH] OR "Hospitalization"[MeSH]	223,999
5	adverse event* OR "drug hypersensitivity"[MeSH] OR "drug toxicity"[MeSH] OR "hyponatremia"[MeSH] OR "seizures"[MeSH] OR "suicide"[MeSH] OR "weight gain"[MeSH] OR "Gastroesophageal Reflux"[MeSH] OR "libido"[MeSH] OR "hepatoxicity"[tw]	228,448
6	"drug interactions"[MeSH]	129,038
7	"Recurrence"[MeSH] OR "remission"[tiab] OR "relapse"[tiab]	248,720
8	#3 AND (#4 OR #5 OR #6 OR #7)	1,668
9	("Randomized Controlled Trial"[Publication Type] OR "Randomized Controlled Trials as Topic"[MeSH]) OR "Single-Blind Method"[MeSH] OR "Double-Blind Method"[MeSH] OR "Random Allocation"[MeSH] OR "Randomized Controlled Trial"[tiab]	464,728
10	#8 AND #9	785

The search strategy and the remaining hits are compiled in **Figure 5.1**.

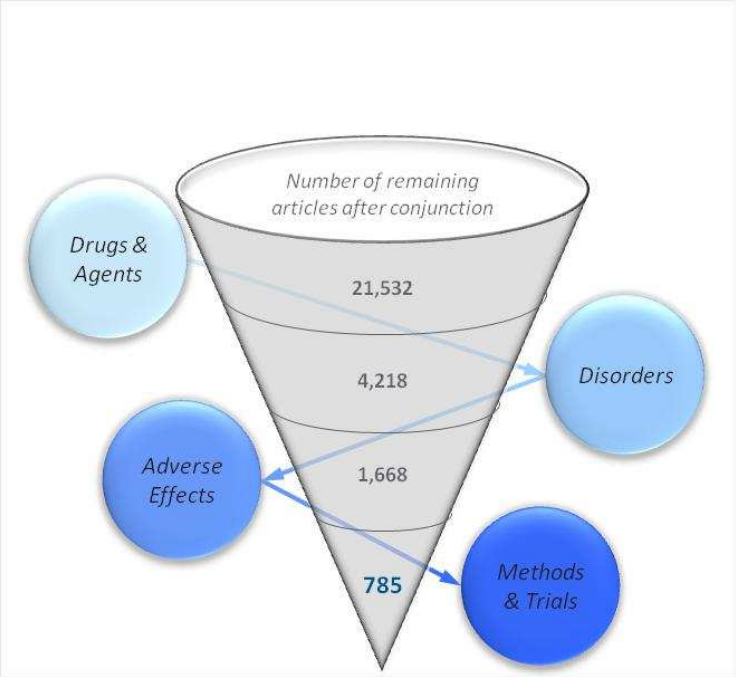

Figure 5.1: Search strategy and remaining hits for Antidepressants for Adult Repression (MMA), data source: PubMed.

Source 2: ClinicalTrials.gov

The search was conducted on 5 December 2011. The search was performed by disjunction of the following search terms (mathematical operations were performed through the Boolean OR function):

- Fluoxetine
- Sertraline
- Paroxetine
- Citalopram
- Fluvoxamine
- Bupropion
- nefazodone
- mirtazapine
- venlafaxine
- escitalopram
- duloxetine
- Trazodone
- desmethylvenlafaxine
- desvenlafaxine

Additionally, the search included the following conditions terms (stated in the search field Conditions):

- "Depressive Disorder"
- "Depressive Disorder, Major"
- "Depression"

Figure 5.2 shows the search mask of ClinicalsTrials.gov including search terms and used search fields. The search resulted in 651 studies.

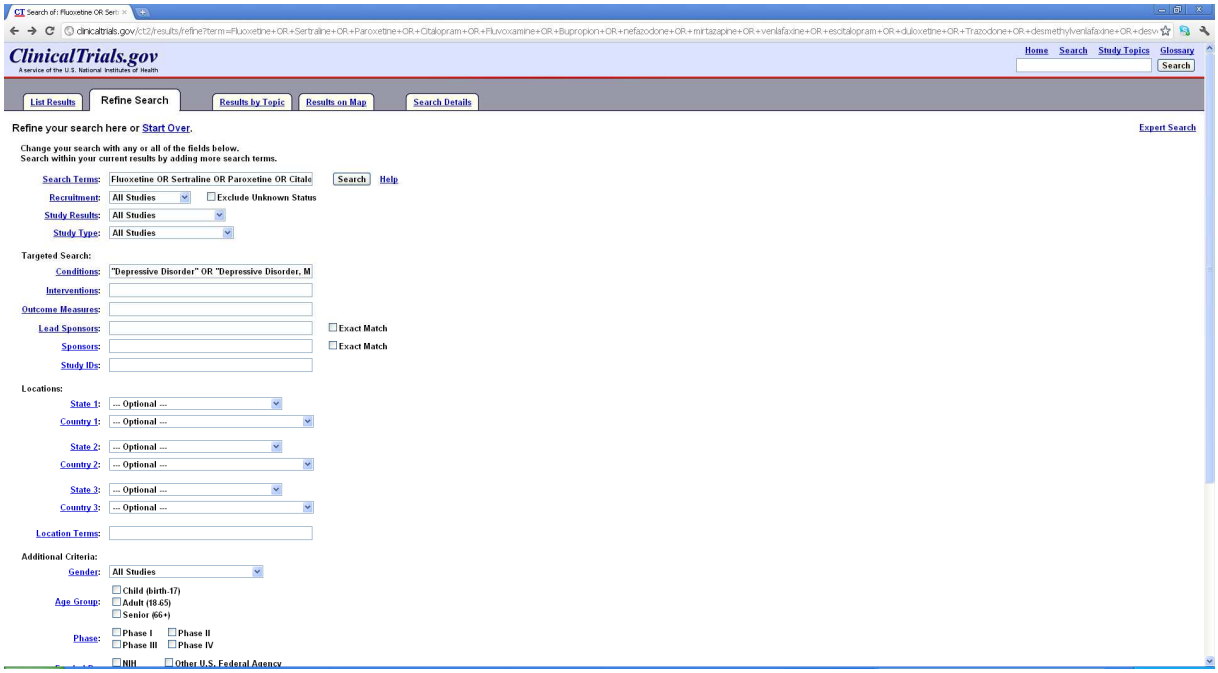

Figure 5.2: Search Mask showing Search Strategy of Antidepressants (MMA) in ClinicalTrials.gov. This screen-shot of the search mask of ClinicalTrials.gov shows the used search fields and any restrictions (no re-strictions).

Source 3: Web of Science

The search was conducted on 24 April 2012. There has been no restriction of the timespan.

Similar to the search in PubMed, the search terms can be divided into blocks. In this case, a combination of all four blocks resulted in an insufficient number of hits. Therefore, the search in Web of Science consisted of the two term blocks drugs and agents, and disorders only. Table 5.2 shows the full search strategy. Search field was the field topic, which in-cludes the title, the abstract and all keywords by the author as well as by the automated tagging of the Web of Science. The search resulted in 3,227 publications in total.

Table 5.2: Search Strategy of Antidepressants (MMA), Source: Web of Science. Search field was the field topic (title, abstract and keywords). The column 'Hits' lists the number of documents retrieved by each search step.

Search Step	Search Strings and Boolean Operations	Hits
1	"antidepressive agent*" OR "anti depressive agent*" OR fluoxetine OR sertraline OR paroxetine OR citalopram OR fluvoxamine OR bupropion OR nefazodone OR mirtazapine OR venlafaxine OR escitalopram OR duloxetine OR trazodone OR o-desmethylvenlafaxine OR desvenlafaxine)	32,396
2	"depressive disorder" OR "dysthymic disorder" OR (depression AND involution*) OR (subsyndrom* AND "depressive disorder"	15,874
3	#1 AND #2	3,227

The search strategy and the remaining hits are shown in **Figure 5.3**.

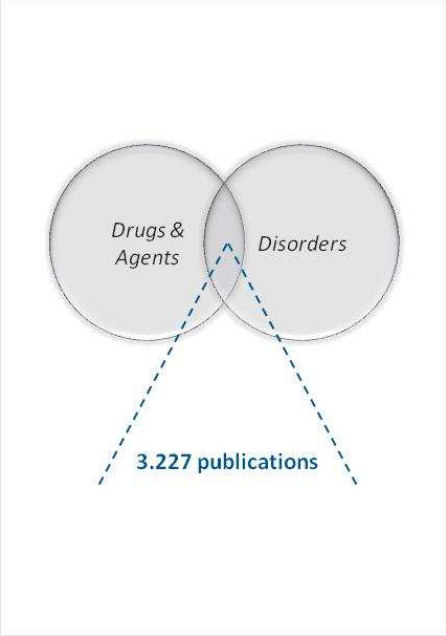

Figure 5.3: Search strategy and remaining hits for Antidepressants for Adult Repression (MMA), data source: Web of Science.

A comparison of the search hits for the MMA topic, given that an analogue search strategy for the searches in both databases is used shows the following picture: When comparing the number of hits for each "search block" (OR-connected block of terms), one finds that the searches in both databases deliver a similar number of hits for the search terms of the first search blocks (in total, 1668 PubMed hits vs. 1428 ISI hits), only the search block for the terms "Randomized Controlled Trial" OR "Single Blind Method" OR "Double Blind Method" OR "Random Allocation" delivers many more PubMed hits than ISI hits (464,728 to 79,635). Consequently, when cutting (AND-connecting) the last block with all the previous ones, 785 PubMed hits remain compared to only 101 ISI hits. In this case, we decided to download the ISI search results previously to the last search block.

5.2.2 Data Set-II: Targeted Immune Modulators (TIM)

Source 1: PubMed

The search was conducted on 2 February 2012. The search strategy shows the same structure like the strategy of data set I. It is again made up of 4 blocks of search terms: drugs and agents, disorders, adverse effects as well as methods and trials. Table 5.3 shows the full search strategy and used search terms. In total, 522 publications were acquired.

Table 5.3: Search Strategy of Target Immune Modulators (TIM), Source: PubMed. In addition to the search terms, the used search field is stated in square brackets after each term. Any restrictions during the search are mentioned accordingly with the label 'limits'. The column 'Hits' lists the number of documents retrieved by each search step.

Search Step	Search Strings and Boolean Operations	Hits
1	"Arthritis, Psoriatic"[MeSH] OR "Crohn Disease"[MeSH] OR "Colitis, Ulcerative"[MeSH] OR plaque psoriasis OR "Arthritis, Juvenile Rheumatoid"[Mesh] OR juvenile idiopathic arthritis	57,492
2	"Arthritis, Rheumatoid"[MeSH] OR ankylosing arthritis *Limits: All Adult: 19+ years*	48,792
3	"adalimumab"[Substance Name] OR humira OR "TNFR-Fc fusion protein"[Substance Name] OR etanercept OR enbrel OR "CDP870"[Substance Name] OR certolizumab OR cimzia OR "infliximab"[Substance Name] OR remicade OR "interleukin 1 receptor antagonist protein"[Substance Name] OR kineret OR anakinra OR "alefacept"[Substance Name] OR amevive OR "abatacept "[Substance Name] OR orencia OR "rituximab"[Substance Name] OR rituxan OR "natalizumab"[Substance Name] OR tysabri	25,314
4	"Treatment Outcome"[Mesh] OR outcome OR efficacy OR effectiveness OR adverse OR safety OR withdrawal* OR harm OR mortality OR morbidity OR function* OR toxicity	5,880,779
5	"Randomized Controlled Trial"[Publication Type] OR "Single-Blind Method"[MeSH] OR "Double-Blind Method"[MeSH] OR "Random Allocation"[MeSH] OR "Case-Control Studies"[MeSH] OR "Cohort Studies"[MeSH] OR "Epidemiologic Studies"[MeSH] OR "Cross-Sectional Studies"[MeSH] OR "Cross-Over Studies"[MeSH] OR "Follow-Up Studies"[MeSH] OR "Multicenter Study "[Publication Type] OR "Evaluation Studies "[Publication Type] OR "Longitudinal Studies"[MeSH] OR "Prospective Studies"[MeSH] OR "Validation Studies"[Publication Type] OR observational studies OR evaluation studies [pt] OR systematic [sb] OR (MEDLINE[Title/Abstract] OR systematic[Title/Abstract] AND review[Title/Abstract] OR meta-analysis[Publication Type])	2,038,605
6	#1 OR #2	103,361
7	#6 AND #3 AND #4 AND #5	1,797
8	#6 AND #3 AND #4 AND #5 *Limits: English, Publication Date from 2008/11/01 to 2010/11/01*	522

Source 2: ClinicalTrials.gov

The search was conducted on 2 February 2013. The search was performed by disjunction of the following search terms (mathematical operations were performed through the Boolean OR function):

- Adalimumab
- Humira
- "TNFR-Fc fusion protein"
- Etanercept
- Enbrel
- "CDP870"
- Certolizumab
- Cimzia
- "infliximab"
- Remicade
- "interleukin 1 receptor antagonist protein"

Additionally, the search included following conditions terms (stated in the search field Conditions):

- Arthritis, Rheumatoid
- ankylosing arthritis
- "Arthritis, Psoriatic"
- "Crohn Disease"
- "Colitis, Ulcerative"
- plaque psoriasis
- "Arthritis, Juvenile Rheumatoid"
- juvenile idiopathic arthritis

Figure 5.4 shows the search mask of ClinicalsTrials.gov including search terms and used search fields. The search resulted in 547 studies.

Source 3: Web of Science

The search was conducted on 24 April 2012. There was no restriction of the timespan.

Similar to the search in PubMed, the search terms can be divided into blocks. In this case, a combination of all four blocks resulted in an insufficient number of hits. Therefore, the search in Web of Science consisted of the two term blocks drugs and agents, as well as disorders only. Table 5.4 gives the full search strategy. Search field was the field topic, which includes the title, the abstract and all keywords by the author as well as by the automated tagging of the Web of Science. The search resulted in 10,713 publications in total.

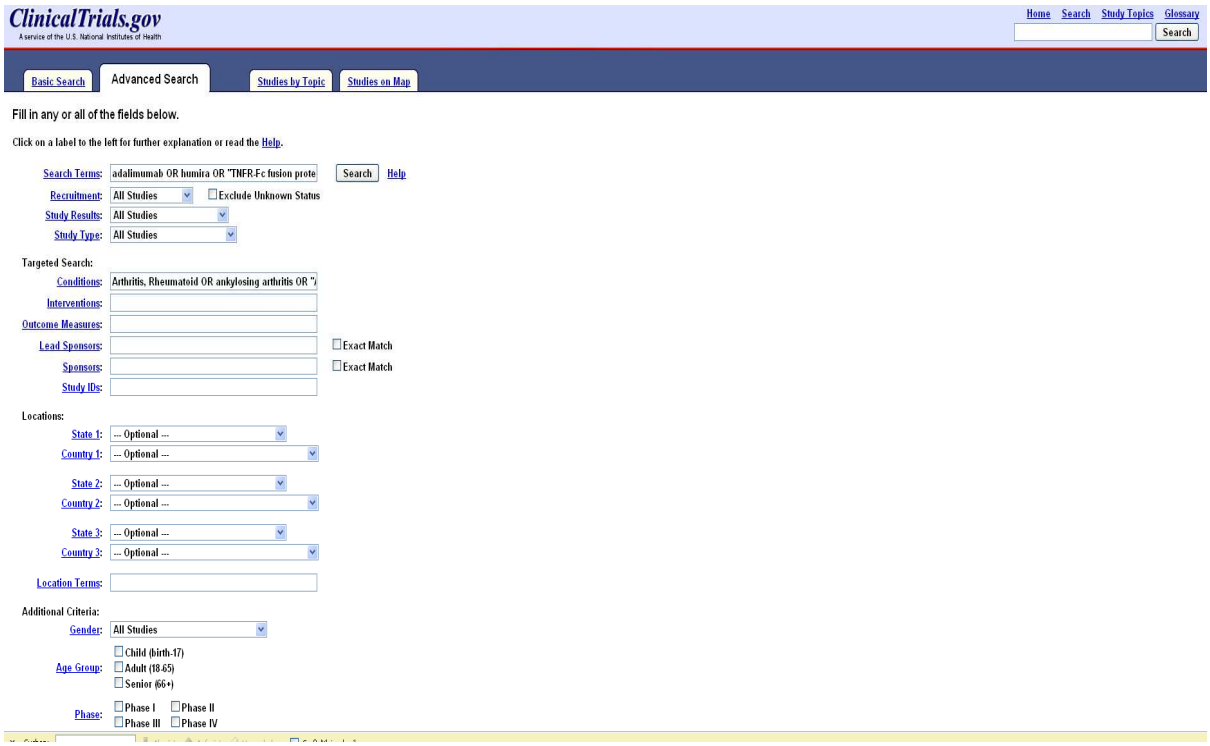

Figure 5.4: Search Mask showing Search Strategy of Target Immune Modulators (TIM) in ClinicalTrials.gov. This screenshot of the search mask of ClinicalTrials.gov shows used search fields and any restrictions (no restrictions).

Table 5.4: Search Strategy of Target Immune Modulators (TIM), Source: Web of Science. Search field was the field *topic* (title, abstract and keywords). The column 'Hits' lists the number of documents retrieved by each search step.

Search Step	Search Strings and Boolean Operations	Hits
1	"psoriatic arthritis" OR "arthritis psoriatica" OR "crohn disease" OR "colitis ulcerosa" OR "ulcerative colitis" OR "plaque psoriasis" OR "juvenile rheumatoide arthritis" OR "juvenile idiopathic arthritis" OR "rheumatoid arthritis" OR (ankylosing AND arthritis)	124,040
2	adalimumab OR humira OR "TNFR-Fc fusion protein" OR etanercept OR enbrel OR "CDP870" OR certolizumab OR cimzia OR infliximab OR remicade OR "interleukin 1 receptor antagonist protein" OR "IL-1RA" OR kineret OR anakinra OR alefacept OR amevive OR abatacept OR orencia OR rituximab OR rituxan OR natalizumab OR tysabri	34,897
3	#1 AND #2	10,713

5.2.3 Data Set-III: Cardiac cardio vascular systems (CCVS)

Source: Web of Science

The search was conducted on 20 April 2013. The time span was restricted to 2005 – actual data.

We restricted the retrieval of publications for this data set to the database Web of Science. The reason is that we identify publications related to a registered clinical trial by the occurrence of an NCT or an ISRCTN number in the abstract and that we do not refer to a PubMed data set used for a systematic review. Table 5.5 shows the full search strategy. Search field was the field *topic*, which includes the title, the abstract and all keywords provided by the author as well as by the automated tagging of the Web of Science. The search resulted in 2,727 publications in total.

Table 5.5: Search Strategy of Cardiac cardio vascular systems (CCVS), Source: Web of Science. Search field was the field *topic* (title, abstract and keywords). The column 'Hits' lists the number of documents retrieved by each search step.

Search step	Search Strings and Boolean Operations	Hits
1	Topic=("single-blind method")	20
2	Topic=("double-blind method")	87
3	Topic=("clinical trial")	57,441
4	Topic=("clinical study")	10,888
5	#4 OR #3 OR #2 OR #1	67,657
6	#4 OR #3 OR #2 OR #1 Refined by: Web of Science Categories=(CARDIAC CARDIOVASCULAR SYSTEMS)	2,727

5.3 Data formats and preparation for further analysis

The PubMed data was delivered in XML files and Endnote files. Data in Endnote files were exported as text readable by BibTechMon™. In case of XML files we had to transform the data in a table like format as BibTechMon™ is working with MS Access Databases. Due to the multi-hierarchical structure of XML data the files had to be transformed. The software scripts that were used to parse and transform the XML data can be found in Section 8.1.

The search results of the ISI Web of Science research were downloaded in a form that can be directly imported into MS ACCESS database tables.

The data available for Task 2.2 of WP2 "Bibliometric analysis of characteristic features distinctive between registered vs. non-registered studies" can be summarized as follows:

- A table of PubMed search results (from DUK's search in December 2011) for set I and II
- A table of WoS search results (from AIT's search in Jan. 2012, and March 2013) for set I, II and III.
- A table of ClinicalTrials.gov search results (from DUK's search in Dec. 2011) for set I and II

6 Results

In this section the available data and its structure are described. In addition, the bibliometric indicators and features based on data availability are discussed.

6.1 Summary of obtained datasets

The search in the different data bases for the three medical topics is summarized in **Table 6.1** and **Figure 6.1.**

Table 6.1: Summary of collected datasets for MMA and TIM.

No.	Database	MMA	TIM	CCVS
(1)	PubMed search	785 (Publ.)	522 (Publ.)	
(2)	WoS overlap with 1 by publication DOI	742 (Publ.)	511 (Publ.)	
(3)	WoS search by topic	3,227 (Publ.)	10,713 (Publ.)	
(4)	Overlap 3 with 2	357 (Publ.)	387 (Publ.)	
(5)	ClinicalTrials.gov	651 (studies)	547 (studies)	
(6)	WoS search			2,727 (Publ.)

The search strategy (1) that was used for the systematic reviews of DUK in the PubMed database delivered for MMA and for TIM 785 and 522 publications, respectively, in PubMed. A more general search in WoS by topic resulted in 3,220 and 10,713 publications, respectively. The overlap of the WoS search by topic with the PubMed search shows only 357 publications.

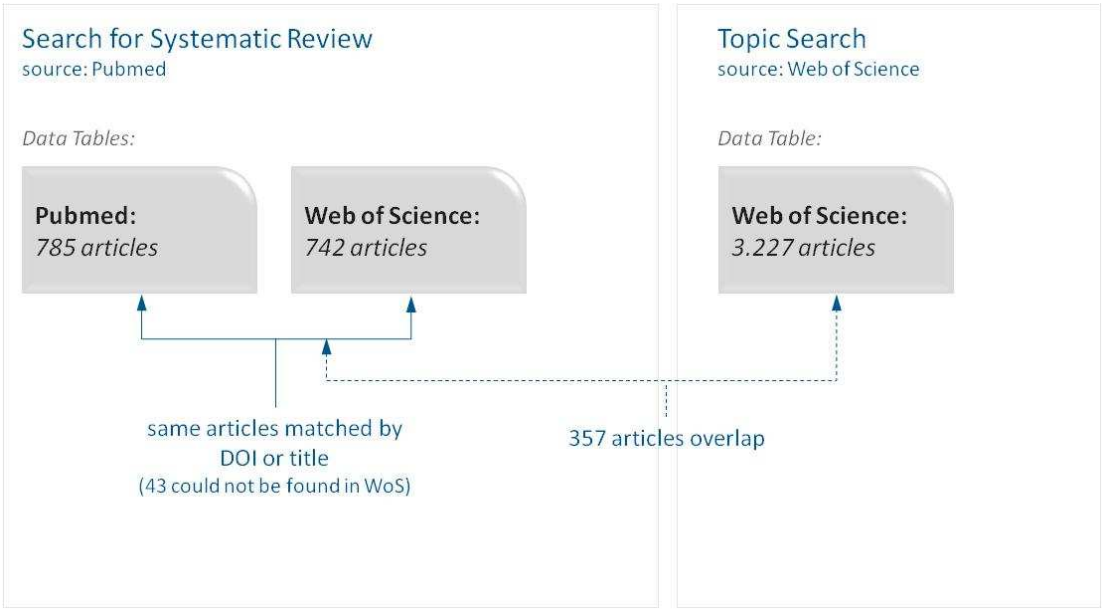

Figure 6.1: Summary of datasets for MMA, data sources: PubMed and Web of Science

The main reason for the high number of relevant results in PubMed can be seen in the better indexing of clinical issues especially in the MeSH thesaurus of PubMed. Web of Science does not have such a classification field. However, it is remarkable that a direct match by the Digital Object Identifier DOI showed an overlap of 742 publications (only 43 or 5.8 % were missing) for the two databases. The high overlap allows a bibliometric analysis with WoS data after an initial search with PubMed. The WoS search by topic can be used to produce a quick overview of other research issues or research in other disciplines going beyond the established search procedures for systematic reviews with a main focus on PubMed. For the topic CCVS we tested a restriction of Web of Science. Since 2005 publications deriving from registered studies have quoted the registration number. We will test if this Information in WoS indicates similar results as in the two other topics.

6.2 Data fields of the Databases PubMed, Web of Science and ClinicalTrials.gov

The results of the comparison of data fields of the three databases are presented in three tables (

Table **6.2**,

Table **6.3**, and Table 6.4). The tables contain the names of the database fields, examples of field entries, numbers of how many of the fields are actually covered, and information on the kind of indicators and bibliometric analyses that can be calculated from the respective database field. Thereby, the presented tables only contain database fields that are certainly or probably relevant for our intended calculations; they do not contain other database fields that are definitely not useful for bibliometric analyses[5].

Each of these tables shows the kind of information the database principally delivers. The tables include

- Names of the database fields,
- Description of contents,
- Indications/suggestions concerning the kind of statistical indicators and bibliometric analyses that can be derived from the data (features and indicators, not all will be used for our analysis)
- Numbers exemplifying the degree of how complete the fields are covered with data, and
- Example entries.

[5] Otherwise, the tables would be much too large, filled with lots of useless information.

Table 6.2: Data Structure of PubMed data set; the number of entries (# MMA) refers to the results of the search for MMA-related publications.

Name	Description	Usability for bibliometric analysis	#Publ MMA	Example
PMID	PubMed Identification Number	Necessary prerequisite	785	16540613
Source	DB: e.g. MEDLINE			MEDLINE
Title	Publication Title	Search, identification, referral	785	Maintenance treatment of major depression in old age.
Mesh Descriptors	MeSH Descriptors (Medical Subject Headings); NLM's controlled vocabulary	Calculation of a network of keywords	785	Depressive Disorder, Major; Female; Male; Double-Blind Method; Aged; Paroxetine; Psychotherapy; Recurrence; Comorbidity; Risk; Humans; Serotonin Uptake Inhibitors; Combined Modality Therapy
Mesh Qualifiers	MeSH Qualifiers (Medical Subject Subheadings)	Calculation of a network of keywords	785	therapy; therapeutic use; therapeutic use; drug therapy; prevention & control
Abstract	Abstract	Search for registration numbers; search for publications with negative results, e.g. "could not reveal", etc.; search for other key terms	776	Elderly patients with major depression, including those having a first episode, are at high risk for recurrence of depression, disability, and death.; We tested the efficacy of …
Number of Ref	Number of References	Potential indicator; however, very scarcely filled out in PubMed (much better in ISI)	26	
Pub Year	Publication Year	Scarcely filled out	141	
Authors	Names of Publication Authors	Search for overlaps between different database sources; Calculation of an author network (based on co-authorship)	784	Reynolds, CF; Dew, MA; Pollock, BG; Mulsant, BH; Frank, E; Miller, MD; Houck, PR; Mazumdar, S; Butters, MA; Stack, JA; Schlernitzauer, MA; Whyte, EM; Gildengers, A; Karp, J; Lenze, E; Szanto, K; Bensasi, S; Kupfer, DJ
Authors Full Name	Full Names of Publication Authors	Search for overlaps between different database sources; Calculation of an author network (based on co-authorship)	784	Reynolds, Charles F; Dew, Mary Amanda; Pollock, Bruce G; Mulsant, Benoit H; Frank, Ellen; Miller, Mark D; Houck, Patricia R; Mazumdar, Sati; Butters, Meryl A; Stack, Jacqueline A; Schlernitzauer, Mary Ann; Whyte, Ellen M; Gildengers, Ariel; Karp, Jordan; Lenze, Eric; Szanto, Katalin; Bensasi, Salem; Kupfer, David J
Affil First Author	Affiliation of First Author (Institution Name and Address)	As only the affiliation of one of the authors is given, NO organization network can be calculated from PubMed (in contrast to ISI)	763	Advanced Center for Intervention and Services Research for Late-Life Mood Disorders, University of Pittsburgh School of Medicine, Western Psychiatric Institute and Clinic, PA 15213, USA. reynoldscf@upmc.edu
Collective-Name			24	STAR*D Study Team
DB Accession Number	CTRegistration Number	Definitely links publication with a specific clinical trial; however only scarcely filled out)	64	NCT00178100
DB Name	CTRegistry Name	Registry name associated with the registry number above	64	ClinicalTrials.gov
Journal	Journal Title Abbreviation		785	N. Engl. J. Med.
Journal Title	Full Journal Title	Statistics about journal policies (in	785	The New England journal of medicine

Name	Description	Usability for bibliometric analysis	#Publ MMA	Example
		connection with knowledge about citation of registration numbers)		
Journal Issue	Journal Issue		738	11
Journal Pub Day	Journal Publication Day		55	16
Journal Pub Month	Journal Publication Month		666	Mar
Journal Pub Year	Journal Publication Year	Statistics about publication years, potential changes in the citation of registration numbers could be observed	746	2006
Journal Volume	Journal Volume		785	354
Journal Page	Journal Pages		785	1130-8
Journal Medline Date			39	
Publication Type	List of Labels Characterizing the Publication Type (Multiple Assignments)	?	785	Research Support, N.I.H., Extramural; Randomized Controlled Trial; Journal Article
Chemical Substances	List of Chemical Substances Names	Network of substances could be calculated	785	Paroxetine; Serotonin Uptake Inhibitors
Chemical Substances Number	List of Chemical Substances Numbers		785	61869-08-7; 0
Ref Type	Associated Reference Type (e.g. reference upon which the article comments, reference containing a comment about the article, reference containing an erratum)		143	CommentIn;
Ref Source	Associated Reference Source		143	ACP J Club. 2006 Jul-Aug;145(1):13; N Engl J Med. 2006 Jun 8;354(23):2505-6; author reply 2505-6; Evid Based Ment Health. 2006 Nov;9(4):101; N Engl J Med. 2006 Jun 8;354(23):2505-6; author reply 2505-6; N Engl J Med. 2006 Mar 16;354(11):1189-90; Evid Base
Ref Source PMID	Associated Reference Source PubMed ID		129	16764055; 16760454; 17076019; 17065298; 16540621; 16813361; 17213124
Journal Info Country	Journal Info Country	Statistics about countries	785	United States
Journal Info Nlm UID	Journal Info NLM UID			0255562
Article ID	List of Article IDs (Multiple Assignments); Article ID values may include the pii (controlled publisher identifier) or doi (Digital Object Identifier).	Search, identification; e.g. to determine the overlap of articles drawn from different databases	785	10.1056/NEJMoa052619; 354/11/1130; 16540613
Article ID Type	List of Corresponding Article ID Types (Multiple Assignments)		785	pii; doi; pubmed

Table 6.3: Data Structure of ISI Web of Science data set; the number of entries (# MMA) refers to the results of the search for MMA-related publications.

Name	Description	Usability for bibliometric analysis	#Publ MMA	Example
UT	Unique Identification Number	Necessary prerequisite	3,227	WOS:000236164200003
AU	Authors	Search for overlaps between different database sources; Calculation of author network (based on co-authorship)	3,227	Trivedi, MH; Fava, M; Wisniewski, SR; Thase, ME; Quitkin, F; Warden, D; Ritz, L; Nierenberg, AA; Lebowitz, BD; Biggs, MM; Luther, JF; Shores-Wilson, K; Rush, AJ
AF	Full Names of Authors	Principally better than "Authors" (less ambiguous); not always filled out	3,227	Trivedi, MH; Fava, M; Wisniewski, SR; Thase, ME; Quitkin, F; Warden, D; Ritz, L; Nierenberg, AA; Lebowitz, BD; Biggs, MM; Luther, JF; Shores-Wilson, K; Rush, AJ
CA	Group Authors		61	STAR D Study Team
SE	Book Series Title		10	
TI	Publication title	Search, identification, referral	3,227	Medication augmentation after the failure of SSRIs for depression.
DT	Document Type		3,227	Article
PT	Publication Type (e.g. J .. journal, B .. Book; S .. Series; C.. Conference)		3,227	J
SO	Source (Publication Name)	Statistics about journal policies (in connection with knowledge about citation of registration numbers)	3,227	NEW ENGLAND JOURNAL OF MEDICINE
J9	29-Character Source Abbreviation		3,227	NEW ENGL J MED
JI	ISO Source Abbreviation		3,222	N. Engl. J. Med.
PY	Publication Year	Statistics about publication years, potential changes in the citation of registration numbers could be observed	3,227	2006
PD	Publication Date		2,803	MAR 23
VL	Volume		3,219	354
IS	Issue		2,848	12
LA	Language		3,227	English
TC	Times Cited	Potential indicator	3,227	305
CT	Conference Title		403	
CY	Conference Year		403	
CL	Conference Location		403	
SP	Conference Sponsors		301	
HO	Conference Host		2	
DE	Author Keywords	Calculation of a network of keywords	1,894	
ID	Keywords Plus®	Calculation of a network of keywords	2,912	STAR-ASTERISK-D; SEQUENCED TREATMENT ALTERNATIVES; ALGORITHM PROJECT; RATIONALE; DISORDER
AB	Abstract	Search for registration numbers; search for publications with negative results, e.g. "could not reveal", etc.; search for other key terms	2,836	BACKGROUND Although clinicians frequently add a second medication to an initial, ineffective antidepressant drug, no randomized controlled trial has compared the efficacy of this approach. …
C1	Affiliation of all Authors (Institution Name and	Organization network can be calculated; statistics about countries	3,089	Univ Texas, SW Med Ctr, Dept Psychiat, Mood Disorder Program & Clin, Dallas, TX

Name	Description	Usability for bibliometric analysis	#Publ MMA	Example
	Address)			75390 USA; Massachusetts Gen Hosp, Clin Psychopharmacol Unit, Boston, MA 02114 USA; Univ Pittsburgh, Grad Sch Publ Hlth, Epidemiol Data Ctr, Pittsburgh, PA USA; Univ Pittsburg
EM	E-Mail Address		2,077	trivedi@utsouthwestern.edu
RP	Reprint Address		2,864	Trivedi, MH (reprint author), Univ Texas, SW Med Ctr, Dept Psychiat, Mood Disorder Program & Clin, 5323 Harry Hines Blvd, Dallas, TX 75390 USA
FU	Funding Agency and Grant Number	?	672	
FX	Funding Text		672	
CR	List of Cited References	Calculation of a network based on co-citations; calculation based on a network of bibliographic coupling	3,016	Rush AJ, 2006, NEW ENGL J MED, V354, P1231; Trivedi MH, 2006, AM J PSYCHIAT, V163, P28; RUSH AJ, 2005, PHYS GUIDE DEPRESSIO, P1; Trivedi MH, 2004, ARCH GEN PSYCHIAT, V61, P669; Rush AJ, 2004, CONTROL CLIN TRIALS, V25, P119, DOI 10.1016/S0197-2456(03)00112
NR	Cited References Count		3,227	18
Z9			3,227	309
PU	Publisher	?	3,227	MASSACHUSETTS MEDICAL SOC
PI	Publisher City		3,227	WALTHAM
PA	Publisher Address		3,227	WALTHAM WOODS CENTER, 860 WINTER ST,, WALTHAM, MA 02451-1413 USA
SN	ISSN	?	3,226	0028-4793
BN	ISBN	?	8	
PN	Part Number		21	
SU	Supplement		345	
SI	Special Issue		24	
BP	Beginning Page		3,186	1243
EP	Ending Page		3,186	1252
AR			41	
DI	Digital Object identifier (DOI)	Search, identification; e.g. to determine the overlap of articles drawn from different databases	2,236	10.1056/NEJMoa052964
PG	Page Count		3,227	10
WC	Web of Science Category	Calculation of subject network	3,227	Medicine, General & Internal
SC	Subject Category	Calculation of subject network	3,227	General & Internal Medicine
GA	Document Delivery Number		3,227	023ZH

Table 6.4: Data Structure of ClinicalTrials.gov data set; the number of entries (# MMA) refers to the results of the search for MMA-related studies.

Name	Description	usability for bibliometric analysis	#Publ MMA	Example
number	Unique ID	Necessary prerequisite	601	442
acronym	Acronym or initials	Search, identification, referral	101	MITT
brief summary text block	Brief summary of the protocol intended for the lay public		651	The purpose of this project is to test whether a new model of collaborative care depression treatment adapted to the needs and preferences of low-income, urban mothers with perinatal depression and to a paediatric clinic setting increases engagement in and adherence to perinatal depression treatment.
brief title	Brief title intended for the lay public	Search, identification, referral	651	Perinatal Depression Treatment in a Paediatric Setting- Pilot Phase
clinical study::ATT1::rank			651	562
completion date	Completion Date	Statistics	573	February 2011
completion date::ATT1::type	Completion Date Type		505	Anticipated
detailed description text	Detailed description including more technical information	Calculation of a keyword network	407	This research project consists of two phases that are designed, in sequence, to adapt standard collaborative care depression treatment to the unique needs of low-income depressed postpartum women and to the paediatric setting where women will receive the treatment. These adaptations, in turn, are designed to …
enrollment	Number of subjects in the trial	Statistics linked with publication behavior	633	15
enrollment::ATT1::type			517	Anticipated
firstreceived date			651	December 8, 2008
firstreceived results			43	
id info nct id	Registration Number	Search, identification, referral	651	NCT00804739
id info org study id		?	651	1R34 MH082141-01
id info secondary id		?	293	
intervention browse mesh term	Intervention mesh term		635	Sertraline
intervention description	Intervention description		425	This is a treatment team approach that allows for outreach. The clinical team will be a nurse practitioner of psychiatry and a social worker. The nurse practitioner will provide either interpersonal psychotherapy, sertraline or both as

Name	Description	usability for bibliometric analysis	#Publ MMA	Example
				indicated.
intervention intervention name	Intervention name		640	Mother-Infant Treatment Team
is fda regulated	Is an intervention that is subject to US Food and Drug Administration regulation included[6]?		467	No
is section 801	If FDA regulated, is it an "applicable clinical trial"[7]?		229	
keyword	Keyword	Calculation of a keyword network	496	Perinatal Depression; Zoloft; Sertraline; IPT; Interpersonal Psychotherapy; Postpartum; Treatment Adherence; Treatment Feasibility; Postpartum Depression; Major Depression; Treatment Engagement
lastchanged date			651	March 8, 2011
location countries country	Country	Country network	601	United States
location facility address city	Location facility address city		601	Rochester
location facility address country	Location facility address country		601	United States
location facility address state	Location facility address state		497	New York
location facility address zip	Location facility address zip		529	14642
location facility name	Location facility name	Search, identification; e.g. can be compared with the name of affiliations of publication authors	576	University of Rochester Medical Center
official title	Official title of study	Search, identification, referral	638	Adapting Collaborative Care Perinatal Depression Treatment to a Pediatric Setting - Pilot Phase
overall official affiliation	Overall official's affiliation		592	University of Rochester

[6] **FDA Regulated Intervention:** Indicates whether a trial includes an intervention subject to US Food and Drug Administration regulation under section 351 of the Public Health Service Act or any of the following sections of the Federal Food, Drug and Cosmetic Act: 505, 510(k), 515, 520(m), and 522.

[7] **Section 801 Clinical Trial:** If a trial includes an FDA regulated intervention, indicates whether this is an "applicable clinical trial" as defined in US Public Law 110-85, Title VIII, Section 801. Briefly, applicable drug trials include controlled clinical investigations, other than Phase I investigations, of a drug or biologic subject to US FDA regulation. Applicable device clinical trials are controlled trials with health outcomes of devices subject to FDA regulation, other than small feasibility studies, and pediatric postmarket surveillance.

Name	Description	usability for bibliometric analysis	#Publ MMA	Example
overall official last name	Person(s) responsible for the overall scientific leadership	Search, identification; e.g. can be compared with the name of authors of publications	596	Linda H Chaudron, MD, MS
overall official role			593	Principal Investigator
overall status	Status of study		651	Active, not recruiting
oversight info authority	Oversight authority (name of national or international health organization with authority over the protocol)	Statistics (e.g. how many of the published registered studies are overseen by this or that oversight authority)	651	United States: Institutional Review Board
oversight info has dmc			424	Yes
phase	Study phase as defined by US FDA (e.g. N/A - without phases, phase 1, phase 2, phase 3, phase 4)	Statistics linked with publication behavior	651	N/A
primary completion date	Is section 801, finishing of final collection of data for the primary outcome		469	December 2010
primary completion date::ATT1::type	Primary completion date type		469	Anticipated
primary outcome description	Primary outcome		99	
responsible party name title	Responsible party name (sponsor that initiates the study, or principal investigator who conducts, or sponsor-investigator: individual who both initiates and conducts the study	Search, identification, e.g. can be compared with the name of authors or organizations (authors' affiliations)	404	Linda Chaudron, MD, MS - Principal Investigator
responsible party organization	Responsible party organization	Search, identification, e.g. can be compared with the name of organizations	403	University of Rochester
results reference citation	Citations to related publications: background and/or results	E.g. search for these publications in PubMed, ISI Web of Science, then checking the text of the publication regarding how the link to the clinical study is given	37	
results reference PMID	PubMed Identifier of related publications	Search, identification	36	
source			651	
sponsors collaborator agency	Sponsor collaborators (other organizations (if any) providing support)	Statistics linked with publication behavior	221	National Institute of Mental Health (NIMH)
sponsors collaborator agency class			221	NIH
sponsors lead sponsor	Sponsor (name of primary organization that over-	Statistics linked with publication behavior	651	University of Rochester

Name	Description	usability for bibliometric analysis	#Publ MMA	Example
agency	sees implementation and is responsible for data analysis)			
sponsors lead sponsor agency class			651	Other
start date	Date the enrollment to the protocol begins		641	January 2009
study design	Study design (e.g. treatment, prevention, diagnostic, screening, supportive care, basic science, health services research, ..)		651	Intervention Model: Single Group Assignment, Masking: Open Label, Primary Purpose: Health Services Research
study type	Study type: interventional, observational, expanded access		651	Interventional
verification date	Date the protocol information was last verified		651	March 2011
why stopped	Brief explanation for suspended, terminated or withdrawn studies		22	

6.3 Bibliometric indicators and features based on a combination of the Data fields

For the bibliometric features and indicators mentioned in Section 4 we need the following information:

- Co-author analysis: standardized author names of all co-publishing authors, ID of the publication
- Co-citation analysis: all standardized citations of the complete reference list, ID of publication
- Bibliographic coupling: ID of Publication and standardized citations of the complete reference list
- Source (Journals): standardized information on Journals, Proceedings, etc.
- Keywords: author's keywords, and indexed keywords from abstracts and titles

and the following indicators:

- Activity indicators: year and ID of the publication
- Times cited: number of citations of the publication

For the application of bibliometric features and the calculation of indicators the available information of the different data sources had to be condensed.

Each calculation needs a unique identifier, the ID, which is available in all databases for each data entry. It will be used for all features and indicators. The title and the abstract are used for content information and automated keyword extraction. Authors are available in PubMed and Web of Science and will be used for the co-author network. The journal title is important. We added information from ICMJE Journals that only accept publications on registered studies.

The ICMJE (International Committee of Medic al Journal Editors) is a working group of general medical journals (no open membership). Its participants meet annually. They work on the Uniform Requirements for Manuscripts (URM) submitted to Biomedical Journals. The section dedicated to *Publishing and Editorial Issues* of the URM include (beside e.g. the *Obligation to Publish Negative Studies*) the *Obligation to Register Clinical Trials,* the essential requirement of which is that "t*he ICMJE member journals require, as a condition of consideration for publication in their journals, registration in a public trials registry*"**.** We will examine whether the related journals have an effect on publications on registered clinical trials.

The publication year is available for time series of publication activity. Standardized cited references is a key feature of WoS and not available in PubMed. They are used for science maps based on co-citation analysis and bibliographic coupling. Additionally, similar publications can be identified. Keywords are used for the retrieval of documents and for a quick naming of clusters in co-citation and bibliographic coupling. PubMed gives information about the registration number of a clinical trial. Web of Science offers the information about the citation count of a publication. It will be used to compare if publications about registered studies have a higher citation rate than others.

Table 6.5: Comparison of data structure of PubMed and Web of Science listing similar and dissimilar fields relevant for the bibliometric analysis.

PubMed		Web of Science		Usability for bibliometric analysis
Data Field	**Description**	**Data Field**	**Description**	
PMID	PubMed Identification Number	UT	Unique Article Identifier	Necessary for document identification, statistics about publications
Abstract	Abstract	AB	Abstract	Information about content, search for registration numbers; search for other key terms
Affil First Author	Affiliation of First Author (Institution Name and Address)	C1	All authors' affiliations	PubMed: affiliation of only one author; WoS: all affiliations, analysis of institutional collaboration possible, countries
Title	Title of the Publication	TI	Document Title	Search, identification, referral
Authors	Names of Publication Authors	AU	Authors	Communities of scientists, network analysis: centralities
Journal	Full Journal Title	SO	Publication Name	Statistics about journal policies (in

PubMed		Web of Science		Usability for bibliometric analysis
Data Field	**Description**	**Data Field**	**Description**	
Title				connection with knowledge about citation of registration numbers)
Journal Pub Year	Journal Publication Year	PY	Year Published	Statistics about publication years, potential changes in the citation of registration numbers could be observed
Ref Type; Ref Source	Associated Reference Type; Associated Reference Source	CR	Cited References	PubMed: references not standardized, WoS:ref. standardized, calculation of co-citations and bibliographic coupling
Chemical Substances; Mesh Descriptors; Mesh Qualifiers	List of Chemical Substances Names; MeSH Descriptors; MeSH Qualifiers (Medical Subject Subheadings)	DE ID	Author Keywords Keywords Plus®	Co-word maps
DB Accession Number; DB Name	CT Registration Number; CT Registry Name	-	-	PubMed: Definitely links publication with a specific clinical trial; (however only scarcely filled out)
-	-	TC	Times Cited	Visibility by citations
-	-	WC	Web of Science Subject Category	Related disciplines

7 Discussion

7.1 Key Findings

Deliverable D2.1 is the basic document for the work to be done in Task 2.2 “Bibliometric analysis of characteristic features distinctive between registered vs. non-registered studies” of WP2, the carrying out of the diverse statistic-bibliometric analyses in the search of features that allow a distinction of the bibliometric profiles of publications based on registered clinical trials vs. those based on non-registered clinical trials. Additionally, we collected and condensed all necessary information to test if relational bibliometric approaches contributes to a broader view in research issues and help to identify more relevant literature for systematic reviews and meta-analysis.

The results are summarized and discussed in Task 2.3 “Characteristic bibliometric features of publication bias and conclusions”, and the identified characteristic features will be tested for their direct or indirect relationships to current prevention measures. This outcome

together with the bibliometric analysis is documented in Deliverable D2.2 of UNCOVER, “Lessons learned from bibliometric analysis and measures against publication bias”.

The content of D2.2, in turn, will contribute (together with results of WP1, WP3, and WP4) to WP6 “Development of recommendations” dedicated to the integration of findings on current and newly proposed measures against publication bias.

8 Appendices

8.1 Processing of XML data

The data provided by DUK was partially delivered in XML format. XML files can consist of many hierarchies and multiple entries for the same element. As our software BibTech-Mon™ cannot import XML data directly, but works perfectly with MS Access, we simply transformed the hierarchical structure of the XML data into a database format. Databases do not allow hierarchies as well as multiple columns with the same name. Therefore the converter compresses the data to one hierarchy and one data field for multiple elements with the same name.

As a first step the XML data was converted into text with a short self-written script in *Ruby*. The script traverses the tree built by the hierarchy with a so called depth-first-search strategy. Elements which contain data are renamed by combining their name with all element names of above hierarchy. This way the hierarchy is restored within the field names of the later database columns and must not be remembered with links. Due to problems with carriage returns within the text, the character string ‘>><<’ marks a new line. After creating the text file with the script carriage returns were replaced by blanks and ‘>><<’ by carriage returns. As a result we got a list of the data with corresponding element names (including the hierarchy) and record number.

As a next step the text list was imported into access and a table was build consisting of one column of each occurring element name. Finally, the table was updated with the data from the list by stringing together data with the same element name separated by semicolon into one data field.

The converter for the different XML files are almost identical and act up to the same principle.

```
########################################
#  converter for pubmed search results  #
#  XML to text-list                     #
#  TIM                                 #
########################################

require 'rubygems'
```

```
require 'builder'
require 'readline'
require 'nokogiri'
require 'rexml/document'
include REXML

output = ''

# creates new file and clears it     w writes, a attaches, r read
# write to the file: file.puts
# write carriage return with \n

xmlfile = File.new('pubmed_all.xml')
xmldoc = Document.new(xmlfile)
n=1


def get_childs(text, child, n, output)
        if child.is_a?(REXML::Text)
        #puts "Text: #{child.to_s.inspect}"
        else if child.has_attributes?
                x=1
                child.attributes.each_attribute {|attr|
                        output= output.concat(">>>< #{n}
        #{text}#{child.name}::ATT#{x}::#{attr.expanded_name}       #{attr.value}")
                        x=x+1
                }
             end #if3

             if child.has_elements?
             child.each do |grandchild|
                newtext="#{text}#{child.name}_"
                get_childs(newtext,grandchild,n, output)
             end #child.each
             else
             output= output.concat(">>>< #{n}       #{text}#{child.name}  #{child.text}")
             end #if2
        end #if1
end #def

XPath.each(xmldoc,'//PubmedArticle/') { |element|
        element.each do |child|
                text=''
                number=1
                get_childs(text,child,n,output)
        end #element.each
        n= n+1
}


fileout = File.open('pubmed_convertedfile.txt', 'w')
fileout.puts output
fileout.close


########################################
#  converter for pubmed search results  #
#  XML to text-list                     #
#  MMA                                  #
########################################

require 'rubygems'
require 'builder'
require 'readline'
require 'nokogiri'
require 'rexml/document'
include REXML

output = ''

# creates new file and clears it     w writes, a attaches, r read
# write to the file: file.puts
# write carriage return with \n

xmlfile = File.new('pubmed_as.xml')
xmldoc = Document.new(xmlfile)
n=1


def get_childs(text, child, n, output)
        if child.is_a?(REXML::Text)
        #puts "Text: #{child.to_s.inspect}"
        else if child.has_attributes?
```

```
                x=1
                child.attributes.each_attribute {|attr|
                        output= output.concat(">>>< #{n}
        #{text}#{child.name}::ATT#{x}::#{attr.expanded_name}       #{attr.value}")
                        x=x+1
                }
             end #if3

             if child.has_elements?
             child.each do |grandchild|
                newtext="#{text}#{child.name}_"
                get_childs(newtext,grandchild,n, output)
             end #child.each
             else
             output= output.concat(">>>< #{n}       #{text}#{child.name}  #{child.text}")
             end #if2
        end #if1
end #def

XPath.each(xmldoc,'//PubmedArticle/') { |element|
        element.each do |child|
                text=''
                number=1
                get_childs(text,child,n,output)
        end #element.each
        n= n+1
}


fileout = File.open('pubmed_convertedfile.txt', 'w')
fileout.puts output
fileout.close


#######################################################
#  converter for Clinicaltrials.gov search results  #
#  XML to text-list                                 #
#  TIM                                              #
#######################################################

require 'rubygems'
require 'builder'
require 'readline'
require 'nokogiri'
require 'rexml/document'
include REXML

output = ''

# creates new file and clears it     w writes, a attaches, r read
# write to the file: file.puts
# write carriage return with \n

xmlfile = File.new('combinedfiles.xml')
xmldoc = Document.new(xmlfile)
n=1


def get_childs(text, child, n, output)
        if child.is_a?(REXML::Text)
        #puts "Text: #{child.to_s.inspect}"
        else if child.has_attributes?
                x=1
                child.attributes.each_attribute {|attr|
                        output= output.concat(">>>< #{n}
        #{text}#{child.name}::ATT#{x}::#{attr.expanded_name}       #{attr.value}")
                        x=x+1
                }
             end #if3

             if child.has_elements?
             child.each do |grandchild|
                newtext="#{text}#{child.name}_"
                get_childs(newtext,grandchild,n, output)
             end #child.each
             else
             output= output.concat(">>>< #{n}       #{text}#{child.name}  #{child.text}")
             end #if2
        end #if1
end #def

XPath.each(xmldoc,'//clinical_study/') { |element|
        if element.has_attributes?
```

```
                x=1
                element.attributes.each_attribute {|attr|
                        output= output.concat(">>>< #{n}     #{ele-
ment.name}::ATT#{x}::#{attr.expanded_name}  #{attr.value}")
                        x=x+1
                }
        end #if
        element.each do |child|
                text=''
                get_childs(text,child,n,output)
        end #element.each
        n= n+1
}

fileout = File.open('clinicaltrials_convertedfile.txt', 'w')
fileout.puts output
fileout.close
```

8.2 List and pertinent information on ICMJE

The ICMJE (International Committee of Medic al Journal Editors) is a small working group of general medical journals (no open membership). Its participants meet annually. They work on the Uniform Requirements for Manuscripts (URM) submitted to Biomedical Journals.

The section dedicated to *Publishing and Editorial Issues* of the URM include (beside e.g. the *Obligation to Publish Negative Studies*) the *Obligation to Register Clinical Trials* the essential requirement of which is that "t*he ICMJE member journals require, as a condition of consideration for publication in their journals, registration in a public trials registry".* This quotation as well as the following ones are taken from *http://www.icmje.org/publishing_10register.html*

Concerning the question where to cite the trial registration number within the text of a publication the ICMJE states the following:

"*The ICMJE recommends that journals publish the trial registration number at the end of the abstract. The ICMJE also recommends that, whenever a registration number is available, authors list this number the first time they use a trial acronym to refer to either the trial they are reporting or to other trials that they mention in the manuscript."*

Concerning the definition of a clinical trial the ICMJE states the following:

"The ICMJE believes that it is important to foster a comprehensive, publicly available database of clinical trials. The ICMJE defines a clinical trial as any research project that prospectively assigns human subjects to intervention or concurrent comparison or control groups to study the cause-and-effect relationship between a medical intervention and a health outcome. Medical interventions include drugs, surgical procedures, devices, behavioral treatments, process-of-care changes, and the like.

Concerning the question of where to register a clinical trial the ICMJE states the following:

The ICMJE does not advocate one particular registry, but its member journals will require authors to register their trial in a registry that meets several criteria. The registry must be accessible to the public at no charge. It must be open to all prospective registrants and managed by a not-for-profit organization. There must be a mechanism to ensure the validity of the registration data, and the registry should be electronically searchable. Trial registration with missing fields or fields that contain uninformative terminology is inadequate."

Currently, the ICMJE accepts registration in the following registries:

- www.anzctr.org.au
- www.clinicaltrials.gov
- www.ISRCTN.org
- www.umin.ac.jp/ctr/index/htm
- www.trialregister.nl
- https://eudract.ema.europa.eu/ (new registrations after June 20, 2011)

In addition to the above registries, starting in June 2007 the ICMJE also accepts registration in any of the primary registries that participate in the WHO International Clinical Trials Portal (http://www.who.int/ictrp/network/primary/en/index.html).

Because it is critical that trial registries are independent of for-profit interests, the ICMJE policy requires registration in a WHO primary registry rather than solely in an associate registry, since for-profit entities manage some associate registries.

Trial registration with missing or uninformative fields for the minimum data elements is inadequate even if the registration is in an acceptable registry.

ICMJE members

Currently, the following journals are members of the ICMJE:

1. Annals of Internal Medicine,
2. British Medical Journal,
3. Canadian Medical Association Journal,
4. Chinese Medical Journal,
5. Croatian Medical Journal,
6. JAMA (Journal of the American Medical Association),
7. Nederlands Tijdschrift voor Geneeskunde (The Dutch Medical Journal),
8. New England Journal of Medicine,
9. New Zealand Medical Journal,
10. Revista Medica de Chile,
11. The Lancet,
12. The Medical Journal of Australia,

13. Tidsskrift for Den Norske Llegeforening (The Journal of the Norwegian Medical Association), and
14. Ugeskrift for Laeger (Journal of Danish Medical Association).

And, the following organizations are participating in the ICMJE:

- U.S. NLM
- World Association of Medical Editors.

However, a large number of non-member journals report that they follow the ICMJE's Uniform Requirements for Manuscripts (URM) Submitted to Biomedical Journals. If a journal follows the ICMJE's Uniform Requirements for Manuscripts Submitted to Biomedical Journals, upon request it can be included on a list of journals that follow ICMJE policy. The list can be found on www.ICMJE.org/journals.html. It includes several hundred journals from all over the world. Out of the listed journals, there may be some that do not follow all of the many recommendations and policies.

The publications of data set I (MMA) obtained from PubMed were published in 142 different journals. We managed to identify 33 of these journals to be listed by the ICMJE. Table 8.1 contains the list of journals of data set I which are listed by the ICMJE.

Table 8.1: List of Journals requested listing by ICMJE (following ICJE recommendations). Journals of publications of data set I: Strategy of Antidepressants (MMA), Source: PubMed

Journal Title
American heart journal
Annals of allergy, asthma & immunology : official
Archives of general psychiatry
Archives of otolaryngology--head & neck surgery
Archives of physical medicine and rehabilitation
Biological psychiatry
BMJ (Clinical research ed.)
Chinese medical journal
Clinical drug investigation
CNS drugs
Comprehensive psychiatry
Current medical research and opinion
Diabetes care
Drugs & aging
Expert opinion on pharmacotherapy
Expert review of neurotherapeutics
International journal of psychiatry in medicine
JAMA : the journal of the American Medical Associa
Journal of general internal medicine
Journal of psychosomatic research
Lancet

Neurology
PharmacoEconomics
Pharmacotherapy
Psychotherapy and psychosomatics
The American journal of cardiology
The American journal of psychiatry
The Annals of pharmacotherapy
The British journal of general practice : the jour
The British journal of psychiatry : the journal of
The Journal of clinical endocrinology and metaboli
The Journal of clinical psychiatry
The New England journal of medicine