3

METHODOLOGY
3.1. Preamble

With the increasing volume of scientific resources on the Web, the citation impact of these electronic resources have been considered in the scholarly communication and new subject field entitled ‘Webometrics’ has been proposed. One of the most important applications of Webometrics is Web citation analysis and comparison with traditional citation (like ISI citation).

Several studies have been analysed the relationship between ISI citations as a scholarly source with Web-based citations (Vaughan and Shaw 1075-87; Zhao 1403-18; Pauly and Stergiou; Kousha and Thelwall 1055-65; Kousha 394-406; Abdoli and Kousha 20-30; Mikki 321-31; Mingers and Lipitakis 613-25). In most cases the Web-based citations correlated with ISI citations and found significant differences in the total numbers of citations and with some exceptions. On the basis of these findings, there have been claims that the Web could be an alternative to the ISI for citation impact calculations (Vaughan and Shaw 1075-87; Kousha and Thelwall 1055-65). Nevertheless, there are differences in the extent to which disciplines publish on the Web and write journal articles (Kling and McKim 890-906; Fry and Talja 20-30) and so more information is needed about discipline differences in online citation counting to confirm or deny these claims. Direct relationship between the ISI citation and Web citation expose the importance of Web citation as a useful resource for evaluation of subject / discipline.
3.2. Statement of Research Problem

The present study explores the commonality between conventional and Web-extracted citation patterns for open access journals in General and Internal Medicine. Keeping in view this, a study on “Comparison of Traditional Citations (ISI Web of Science and Scopus) with Web Citations (Google Scholar) of articles of select Medical journals” has been undertaken.

3.3. Significance of the Study

There have been several studies used Web citations to determine the potential use of Web resources for impact assessment and research evaluation (Kousha and Thelwall 1055-65; Kousha 394-406; Vaughan and Shaw 1313-22; Vaughan and Shaw 1075-87), but there is less evidence on online impact measures and Web extracted citation indicators in Medicine and despite all the reviews, no previous studies used extensive data collection method and analysis in Medicine to compare conventional citations and web-extracted citations.

The approach of comparing conventional citations with Web-extracted citations helps us to understand the extent of citations available on the Web and their functionality as a useful or trivial source for Medicine research evaluation in general, and in particular the correlation tests used as an indirect approach to assess the extent of the relationship between traditional and Web-based citation patterns. The results can be considered as additional evidences to show the commonality between traditional and Web -extracted citations in Medicine research. The result, with help of comparing traditional and Web-based citation patterns, can be explored whether
the Web citation extraction techniques and tools could be considered as substitute for the conventional citations.

This proposed research is one of the few attempts to study the characteristics of ‘unique’ citing sources from Google Scholar, to explore the type, language and publication year of the sources of Google Scholar unique citations to articles in ISI-indexed open access journals in Medicine. It is, thus, an exploratory study of importance to identify the causes of differences between conventional and Web-extracted citation patterns and to shed light on the role of non-journal documents such as conference papers, e-prints, research reports and dissertations in formal scholarly communication in the Medical research. The results of this study can also be helpful for researchers and students to answer the question whether researchers and students, especially those who have no access to fee-based citation indexes (i.e., ISI Web of Science and Scopus) can use Google Scholar for locating scholarly information?

More knowledge about open access scholarly publishing on the Web would be helpful for citation data mining and the development of Web-based citation indexes. This research may shed light on to design and development of scientific Web mining tools (e.g. Web-based citation indexing).

3.4. Operational Definitions

The definitions of the key terms used in this research are listed below:
• Citation Analysis

Citation Analysis is the analysis of data derived from references cited in footnotes or bibliographies of scholarly publications and is the best known of Bibliometric approaches, which has been used as a powerful and popular method of measuring and mapping the intellectual impact and visibility of scientists, projects, journals, disciplines, and nations.

• Citation Counts

A Citation count is the number of times an article is cited by other articles.

• Cited document

A Cited document refers to a document that receives citations from other documents.

• Citing document

A Citing document is a document that citing other documents.

• Correlation Test

Correlation is a measure of the relation between two or more variables. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation while a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of correlation. The most widely-used type of correlation coefficient is Pearson r and Spearman Correlation Coefficient.

• Data Mining
Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarising it into useful information.

- **Directory of Open Access Journals**

  This directory is one of the comprehensive and popular subject directories that are indexing the full-text scholarly journals. Now, the *Directory of Open Access Journals* (DOAJ) indexes more than 8000 journals full-text in all scientific subjects (About DOAJ) that nearly 400 journals can be found with a Journal Impact Factor.

- **EndNote**

  EndNote is a commercial reference management software package used to manage bibliographies and references / Citations. It is developed by Thomson Reuters.

- **General and Internal Medicine**

  The branch of Medicine that deals with the diagnosis and (nonsurgical) treatment of diseases of the internal organs (especially in adults).

- **Google Scholar**

  *Google Scholar* <http://scholar.google.com> is a free search engine aimed at finding scholarly information on the Web, and it has been available in its beta version since November 2004. *Google Scholar* is the largest new multi-disciplinary tool that can be used for accessing and measuring ‘hidden’ citations, especially from
non-journal Web documents which are not traced by conventional bibliographic and citation database, such as conference papers, theses, books, research/technical reports, and preprint repositories.

- **Impact Factor**

  It is a measure of the frequency with which the average article in a journal has been cited in a given period of time.

- **ISI Web of Science**

  *Web of Science* is an online citation index provided by Thomson Reuters. Its indexing coverage start from the year 1900 to the present, that covers over 40 million records from 256 disciplines of the Sciences, the Social Sciences and the Arts and Humanities (About *Web of Science*).

- **Journal Citation Reports (JCR)**

  *Journal Citation Reports (JCR)* is an annual publication by the Healthcare & Science Division of Thomson Reuters. It integrated with the *Web of Knowledge* and is accessed from the *Web of Science*. It provides information about academic journals in the Sciences and Social Sciences, including impact factors. It is the trusted and time-tested journal evaluation resource that provides quantifiable, statistical information based citation data.

- **Microsoft Excel**
Microsoft Excel is a spreadsheet application written and distributed by Microsoft. Microsoft Excel provides calculation, graphing tools, pivot tables, macro programming and data analysis.

- **Nonparametric Test**

  The word nonparametric means that there are no requirements on parameters such as standard deviation. Nonparametric Test is also called distribution free tests because they do not require that the data be normally distributed.

- **Open Access Journals**

  Open access journals are scholarly journals that are available online to the reader without any financial, legal, or technical barriers other than those inseparable from gaining access to the Internet itself.

- **Overlapping**

  The amount of records from one database that found in another database.

- **Qualitative Analysis**

  Qualitative Analysis is focused on differences in quality rather than differences in quantity. Results are in words or pictures rather than numbers. Qualitative studies usually have fewer participants than quantitative studies because the depth of the data collection does not allow for large numbers of participants.

- **Quantitative Analysis**
Quantitative Analysis focuses on numbers or quantities. Quantitative studies have results that are based on numeric analysis and statistics. Often, these studies have many participants.

- **RIS format**

  RIS is a standardised tag format developed by Research Information Systems, Incorporated to enable to exchange data. RIS files can be exported from reference software such as EndNote and Reference Manager.

- **Scatter Plot**

  Scatter Plot is a graph of plotted points that show the relationship between two sets of data. A Scatter Plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient.

- **Scopus**

  *Scopus* is the abstracting and citation database of peer-reviewed literature with smart tools that track, analyse and visualise research. Coverage of Scopus goes back to 1966 (1996 for citations). It covers nearly 20,500 journal titles from over 5,000 international publishers, including coverage of 16,500 peer-reviewed journals in the Scientific, Technical, Medical, and Social Sciences (including Arts and Humanities). Subject areas covered in Scopus are Chemistry, Physics, Mathematics, and Engineering, Life and Health Sciences, Social Sciences, Psychology, and
Economics, Biological, Agricultural, and Environmental Sciences, and General Sciences.

- **Scholarly Communication**

  Scholarly communication is the system through which research and other scholarly writings are created, evaluated for quality, disseminated to the scholarly community, and preserved for future use. The system includes both formal means of communication such as publication in peer-reviewed journals, and informal channels such as electronic Listservs.

- **Spearman Correlation Coefficient**

  As the nonparametric counterpart of the Pearson Correlation Coefficient, the Spearman Correlation Coefficient can be used to determine whether two variables are related when the requirements for the Pearson Correlation Coefficient test are not met. The Spearman Correlation Coefficient, also referred to as Spearman’s rho, is symbolised by $r_s$, as opposed to $r_p$ for Pearson correlation coefficient.

- **SPSS**

  SPSS stands for Statistical Package for Social Sciences and is a comprehensive system for analysing data, which consist of a set of tools for data entry, data management, statistical analysis and presentations.

- **Unique Citation**
Unique citations refer to those citations from one of the databases which did not occur in the other databases.

- **Venn diagram**

  A type of graphic organiser that uses overlapping circles to visually represent the commonalities amongst sets of information. It can use two or more circles; the spaces that don't overlap contain unique details and the space where the circles overlap contains information that is the same for each circle.

- **Web extracted / based citations**

  Citations that taken from the Web as a source.

- **Zotero**

  *Zotero* is free and open-source reference management software to manage bibliographic data and related research materials. It has been used to download citations from *Google Scholar*.

### 3.5. Research Questions

The following research questions were formulated to analyse the quantitative and qualitative similarities between traditional citation pattern (*ISI Web of Science* and *Scopus*) and citations taken from the Web (*Google Scholar*).

1. Are there any significant correlation between traditional citation pattern and citations taken from the Web to articles of select open access ISI-indexed Medical journals?
(a). Is there a correlation between *ISI Web of Science* citation counts and *Google Scholar* citation counts?

(b). Is there a correlation between *Scopus* citation counts and *Google Scholar* citation counts?

2. Are they any overlap between these citations? If so, what is the overlap percentage between traditional citations (*ISI Web of Science* and *Scopus*) and *Google Scholar* citations?

3. What are the characteristics of the unique citations of *Google Scholar*, *ISI Web of Science* and *Scopus*?

   (a). What is the document type (i.e., journal articles, conference/workshops papers, research/technical reports, thesis/dissertations, books/book chapters, e-print archives and preprints) distribution of *Google Scholar* citing sources?

   (b). What medium of language (English/other languages) and publication year do *Google Scholar* unique citations have?

   (c). What are the characteristics of sources of *ISI Web of Science* citations which do not overlap with *Google Scholar* citations?

   (d). What are the characteristics of sources of *Scopus* citations which do not overlap with *Google Scholar* citations?
3.6. Objectives

This study attempts to analyse the quantitative and qualitative similarities between traditional citation pattern (ISI Web of Science and Scopus) and citations taken from the Web (Google Scholar) that targeting articles of open access ISI - indexed journals from Medicine published in the year 2007.

The specific objectives are:

1. To analyse the quantitative (citation count) similarities between traditional citation pattern (ISI Web of Science and Scopus) and citations taken from the Web (Google Scholar) in Medicine;

2. To study the correlation between traditional citation pattern and citations taken from the Web of articles of select open access ISI-indexed Medical journals;

3. To assess the overlap percentage between ISI Web of Science and Scopus citations, and Google Scholar citations of articles of select open access ISI – indexed Medical journals,

4. To identify the characteristics of the Google Scholar unique citations (type, language and publication year) to articles of select Open Access ISI – indexed Medical journals; and

5. To identify the characteristics of ISI Web of Science and Scopus unique citations to articles of select open access ISI- indexed Medical journals.
3.7. Scope and Coverage

In order to achieve the objectives of this research, the scope and coverage of the research was set out.

- **Area of study**: General and Internal Medicine.
- **Period**: ISI-indexed OA journals published in the year 2007.
- **Publication Type**: Only the research articles were considered.
- **Language**: The articles that are published in English language were only selected.
- **Sources**: Directory of Open Access Journals (DOAJ), Journal Citation Report (JCR), ISI Web of Science, Scopus and Google Scholar.

3.8. Justification for Scope and Coverage

Citation analysis is an important tool used to trace scholarly research and to monitor a subject trend (Bauer and Backkalbasi). This study being one such kind has its scope and coverage (Section 3.7) confined on certain justifications, which are as follows:

- Since, the research is comparing conventional citations and Web-extracted citations, the following sources have been chosen:
  - *Journal Citation Report (JCR)*: It is a source for identifying the world’s most highly-cited, peer reviewed journals. It is covering 10,196 journal listings in 238 disciplines of Science and Social Sciences.
- **Directory of Open Access Journals (DOAJ):** A directory covering majority of the OA journals published in almost all disciplines. It is one of the frequently using directories.

- **ISI Web of Science:** It is conventional/traditional citation indexing source covering more than 12,000 journals in all disciplines and provides access to the *Science Citation Index, Social Sciences Citation Index* and *Arts & Humanities Citation Index*.

- **Scopus:** Contains 41 million records, nearly 20,500 journal titles from 5,000 publishers worldwide, includes over 3 million conference papers and covers Science, Life Science, Medicine, Engineering etc. It is also one of the comprehensive sources from *Elsevier*, considered as an important source and used predominantly.

- **Google Scholar:** Contains citation information from many publishers, including the coverage of non-journal documents, such as conference papers, theses, books, research/technical reports, and preprint repositories.

  - Considering the features of availability and accessibility refereed or editor-reviewed open access journals in Medicine have been chosen for data gathering.

  - Open access journals being the major and nascent carrier of scientific communication and have rapidly become a global environment for scholarly communication. A significant portion of the scientific literature on Medicine can now be found appearing in the peer-
reviewed open access journals. Hence, the open access journals have been only chosen.

- In order to allow significant time window in which to attract the citations (minimum period was set out is 3 years) open access journals that had been published in 2007 chosen for this study.
- English language has been chosen for this study because it is the carrier language of scientific communication and plays a significant role on scholarly communication.

3.9. Limitations of the Study

The current study only Web citations that taken from the Google Scholar, for comparing Web citations with traditional citations, was used but other Web citation methods like Web/URL were not used.

3.10. Hypotheses

Based on the objectives of the study, the following hypotheses were formulated:

1. There is a significant relationship between ISI Web of Science citations counts and citations taken from the Google Scholar.
2. There is a significant relationship between Scopus citations counts and citations taken from the Google Scholar.
3. There is a significant percentage of overlapping occur between ISI Web of Science citations and Google Scholar citations.
4. There is a significant percentage of overlapping occur between Scopus citations and Google Scholar citations.

3.11. Methodology

The following methodology has been adopted to achieve the objectives of this research. The steps are as follows:

Step – 1: Selection of Disciplines and Journals

1.1. For achieving the objectives of this research and as per the scope set out for the study, as the first step, the Journal Citation Reports (JCR) was searched to find out the journals with impact factor (IP) in General and Internal Medicine subject area.

1.2. The result of this search yielded 153 journals in the chosen disciplines.

1.3. Each of the journals under the subject category of General and Internal Medicine in JCR was searched in the Directory of Open Access Journals (DOAJ) in order to find out the OA journals that covering the disciplines. This process yielded 23 OA journals that indexed in ISI Web of Science. These 23 OA journals gave 2,082 research articles in English language that are published in the year 2007.

Step – 2: Selection of Databases

2.1. As per the scope of this study, the following databases have been used:

- ISI Web of Science (http://www.isiwebofknowledge.com/);
- "Scopus" ([http://www.scopus.com](http://www.scopus.com)); and

- "Google Scholar" ([http://scholar.google.co.in/](http://scholar.google.co.in)).

**Step - 3: Data Collection - Citation Counts**

3.1. Title of 2,082 articles appeared in 23 journals were exported from *ISI Web of Science* to EndNote X5 for further analysis.

3.2. Collection of citation counts: Each of 2,082 research articles was searched in the three databases (*ISI Web of Science*; *Scopus*; and *Google Scholar*) in order to find out the citation counts. The processes are as follows:

3.2.1. The number (for quantitative study) and characteristics (for qualitative study) of citing sources of each article were identified using the option “times cited” in *ISI Web of Science* and exported to EndNote X5. As a result, 15,845 records (citations) were yielded in this process.

3.2.2. *Scopus* was searched by using title of each article, and the number of citations was recorded using the option “Cited By” field of the respective bibliographic record. Since *Scopus* does not provide the Z39.50 data extraction, to export data in EndNote the citing sources for each article have been downloaded as a “RIS” file format and exported the same to EndNote X5. 19,015 records were exported to EndNote.

3.2.3. For *Google Scholar* citation counts, the titles of all 2,082 articles searched as phrase searches in the *Google Scholar* search page and the number of citation counts recorded for each article by clicking the “cited
by” option available below of each retrieved record. *Google Scholar* also does not support the Z39.50 protocol, *Zotero* (open source software) was used for download the citing sources of selected articles. The citing sources for each article have been saved as a “RIS” file format in order to export to EndNote X5. In this process 28,040 records were obtained.

**Step - 4: Data Analysis**

Citations of the research articles were made available for data analysis. The processes for preparation of data for analysis are as follows:

4.1. The exported citations for each article (total 2,082) were formed working database by creating separate EndNote Library.

4.2. The “Master EndNote Library” was formed by importing all EndNote Library files (2,082) for each journal. This yielded 23 “Master EndNote Library”. This process was adopted for *Scopus* and *Google Scholar* also.

4.3. Identification of citation overlapping and unique citations: An examination has been carried out with all citations (62,900) appeared in 23 journals to determine the citation overlapping and unique citations. The processes involved in this are as follows:

4.3.1. Using overlapping algorithm which was developed for determining overlap between two databases. First, the overlap between *Google Scholar* and *ISI Web of Science* (GS ∩ ISI), the overlap between *Google Scholar* and *Scopus* (GS ∩ Scopus) were determined; secondly the
overlap between *ISI Web of Science* and *Scopus* (*ISI ∩ Scopus*) was done. Finally, the overlaps between *Google Scholar*, *ISI Web of Science* and *Scopus* (*GS ∩ ISI∩ Scopus*) were identified. These processes of identification of overlapping citations between databases were carried out in the following ways:

4.3.1.1. The *Google Scholar* and *ISI Web of Science* citing references Endnote files of each journal were merged to explore the overlap among these two databases (*GS ∩ ISI*) through “EndNote Duplication Check”.

4.3.1.2. Similarly, the overlap between *Google Scholar* and *Scopus* (*GS ∩ Scopus*) and also overlap between *ISI Web of Science* and *Scopus* (*ISI ∩ Scopus*) were identified.

4.3.1.3. To find out the overlapping between three databases (*ISI ∩ GS ∩ Scopus*), the *Google Scholar*, *ISI Web of Science* and *Scopus* citing references Endnote files for each journal were merged and duplicates between them were identified.

4.3.2. By omitting the overlapped citations records between databases the unique citations of each databases were identified.

4.3.2.1. *Google Scholar* Unique Citations: The process of omitting overlapped citations records between *Google Scholar* and *ISI Web of Science* for 23 journals led to identify the *Google Scholar* unique citations which does not occur in the *ISI Web of Science* citations and
ISI Web of Science unique citations which does not occur in the Google Scholar citations. The Google Scholar unique citations exported to 23 EndNote Library files. Similarly, by omitting overlapped citations records between Google Scholar and Scopus for 23 journals the unique citations were identified for Google Scholar and Scopus. The Google Scholar unique citations exported to 23 EndNote Library files. In order to find out Google Scholar unique citations, which do not occur within the two other databases (ISI Web of Science and Scopus) for the current study, the EndNote Library files of unique citations from Google Scholar that found above were merged. The compilation was done by importing these EndNote Library files into the “Master EndNote Library” and deleted the duplicate records. This exercise, after checking duplication, yielded a total of 18,004 Google Scholar unique citations.

4.3.2.2. ISI Web of Science Unique Citations: The ISI unique citations after omitting the overlapped records, exported to 23 EndNote Library files. And also by omitting overlapped citations records between ISI Web of Science and Scopus for 23 journals (that found in step 4.3.1.2), the ISI unique citations which do not occur within the Scopus and the Scopus unique citations which do not occur within the ISI Web of Science were identified. The unique citations of ISI Web of Science which does not occur in the Google Scholar, and the unique citations of ISI Web of Science which does not occur in the Scopus were merged in a single “Master library files” for each journals. A total of 6,322 unique citations of ISI Web of Science were received.
4.3.2.3. Scopus Unique Citation: Similarly the Scopus unique citations were identified and yielded 7,168 records.

4.4. By using “EndNote Duplication Check” option in the EndNote X5, the duplicate records were identified from 23 EndNote Library unique citations files for each database. For this purpose, title, author, year, page number and publisher fields were used.

4.5. The citations in the working databases of EndNote X5 files were exported to Microsoft Excel for further analysis.

4.6. The characteristics study, descriptive analysis, etc. were carried out and presented in the next chapter (Chapter - 4).

3.12. Conspectus

Based on the objectives of this study and with defined methodology, a data set arrived for data analysis. Using appropriate statistical tools and techniques results are yielded. The results and interpretations are presented in the Chapter 4.