SensPrecOptimizer

User Manual

Title A software tool for combining search queries to design efficient search strategies

Version 2.0.X

Depends None

Author Mohsen Mesgarpour <mohsen.mesgarpour@gmail.com>

Contributors Bita Mesgarpour, Harald Herkner

Date 2013-10-28

Description Development of Highly Sensitive Search Strategies, Sensitivity and precision analysis for “OR” combination of search queries

Supported OSs Windows (6.0 or above) or Mac (10.4 or above) or Linux (Fedora 17 or above)

License Apache License 2.0

Input and Output Encoding UTF-8

3rd Party Software Needs MS Office, Apache OpenOffice

Repository Google Code

Available for download at: https://code.google.com/p/sens-prec-optimizer/

Table of Content

Description

SensPrecOptimizer is a freeware software to implement profiling of Highly Sensitive Search Strategies (HSSS) by combining the search queries. HSSS refers to the two strategies recommended by the Cochrane Handbook for Systematic Reviews of Interventions: a sensitivity-maximizing version and a sensitivity- and precision-maximizing version [1]. This software may also be used for developing search filters or hedges.

1. Download and Setup

1.1. Hardware and Software Requirements

The minimum required hardware settings are 1GB of RAM and 1GB of free memory storage, and the desired requirements are 2GBs of RAM, 3GBs of free memory storage and a multi-core processor. SensPrecOptimizer needs 80 MB free storage for installation.

The supported Operating Systems (OS) are:

- MS Windows OS (6.0 and above)
- Apple Mac OS (10.4 and above)
- Linux OS (Fedora 17 and above)

1.2. Setup

Please follow the below instructions to install the SensPrecOptimizer. If your OS is not listed in below, the SensPrecOptimizer may not be compatible.

1.2.1. MS Windows OS (6.0 and above)

1. Download and decompress the Windows' version of the package;
2. Open the directory;
3. Double-click on the SensPrecOptimizer shortcut to execute the programme.

1.2.2. Apple Mac OS (10.4 and above)

1. Download and decompress the Mac's version of the package;
2. Download and install the Xcode from the Mac App Store (it is a must for the Alpha version)
3. Download and install the QT Creator (it is a must for the Alpha version)
4. Double-click on the SensPrecOptimizer shortcut to execute the programme.

1.2.3. Linux OS (Fedora 17 and above)

The source code is ready for Linux environment; however, it needs additional recourses for test and compile.

2. Graphical User Interface

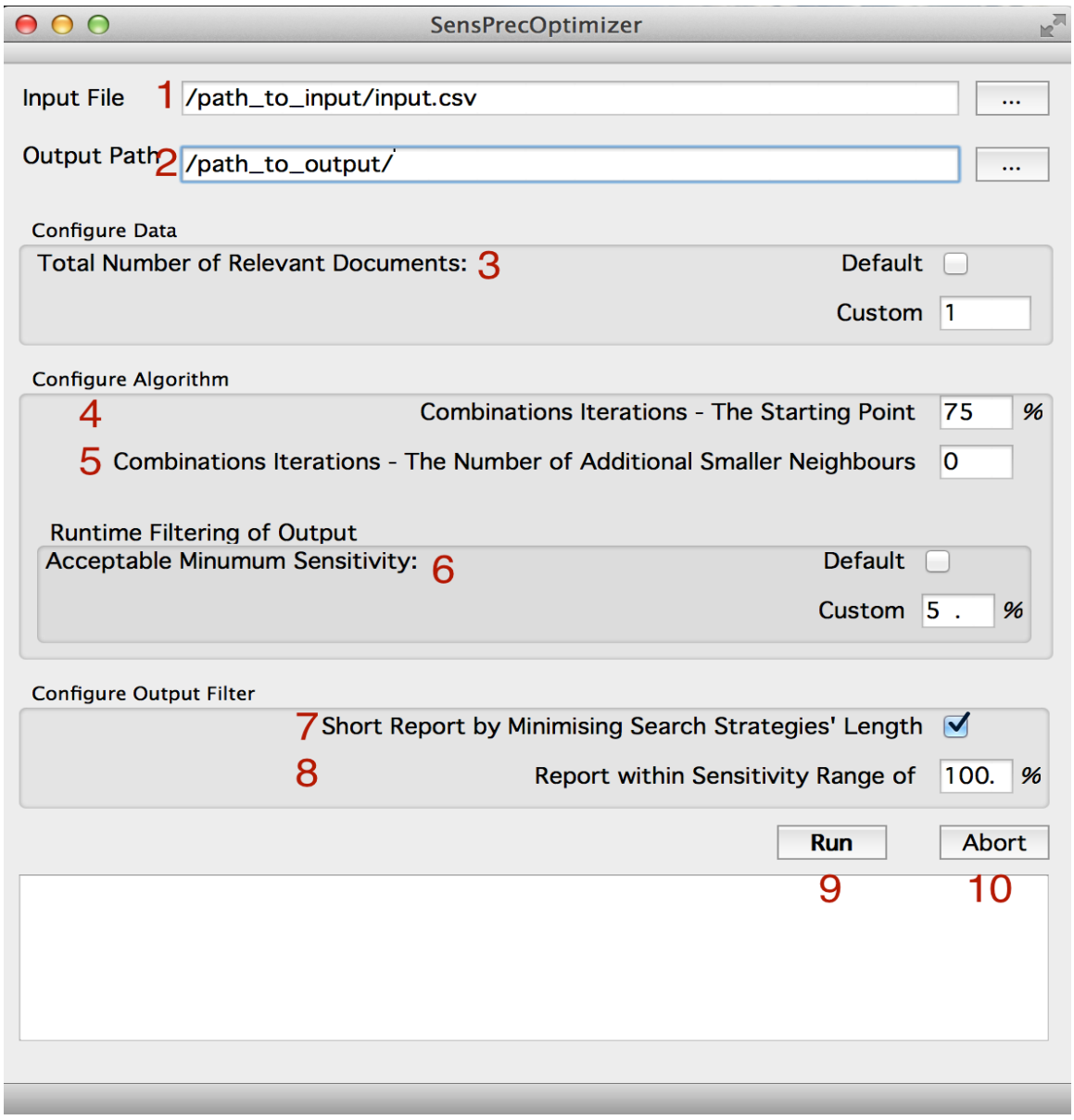


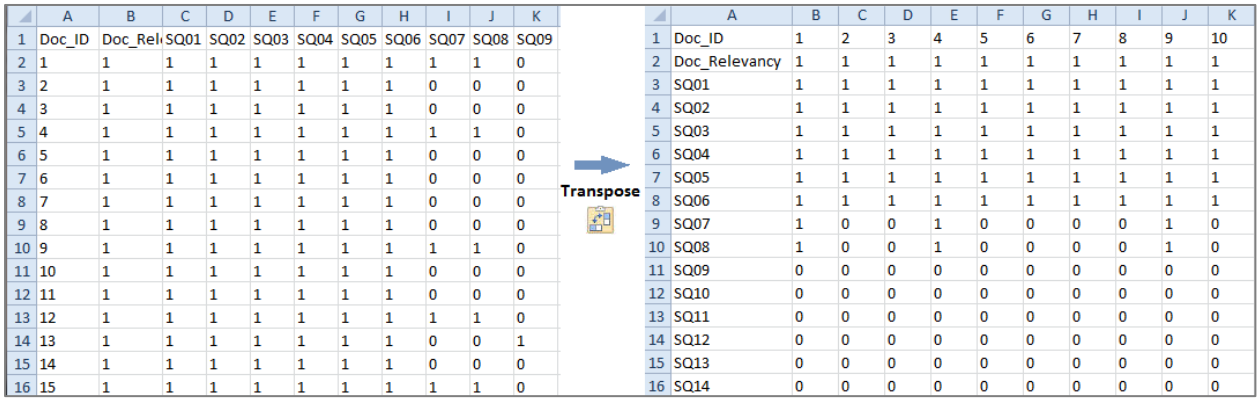
Fig. 1. A screen shot of the SensPrecOptimizer's User Interface

The Graphical User Interface (GUI) consists of (Fig. 1) the following entities:

1. Specify the absolute path to the input file.
2. Specify the absolute path to the output.
3. Set the total number of relevant documents. By default, it is summing up the number of relevant documents (the second row) in the input file.
4. Set a combination length (percentage) for the algorithm to start with. It is recommended to be set to a value below 40% or above 60%.
5. Set the number of additional smaller combinations to be calculated by the algorithm. By default, the algorithm stops when the sensitivity drops below the maximum sensitivity in the current iteration. It is recommended to be set to a value between 0 and 5.
6. In order to optimize the memory usage, specify a narrow sensitivity range for the algorithm. By default, it is set to the maximum sensitivity value of the search queries. It is especially important to set a narrow value for search queries that are longer than 21.
7. It filters down longer search strategies with similar sensitivity and precision, resulted in the shortest strategy with the highest sensitivity and precision.
8. In order to minimize the outputted strategies for long search strategies (more than 21 queries), specify the acceptable sensitivity range for output filter. The range is the distance from the maximum sensitivity.
9. Run the algorithm.
10. Abort the execution of the algorithm.

3. Input File

To construct a compatible input file, an Excel or Access file with a matrix of search queries and documents (or records) is demanded. By searching a query (one column for each query), a set of documents (one row for each document) will retrieve which should be recorded with all necessary characteristics and an assigned ID code. The relevancy of each document should be identified and recorded in a separate column (preferably the second column). Therefore, the ID, relevancy and queries form the consecutive columns in the input file (Fig. 2). The retrieval status of the records retrieved by a search query should be specified as ‘‘one’’ and the rest of records as ‘‘zero’’ in a column (the values must be numerical).



	A	B	C	D	E	F	G	H	I	J	K
1	Doc_ID	Doc_Rel	SQ01	SQ02	SQ03	SQ04	SQ05	SQ06	SQ07	SQ08	SQ09
2	1	1	1	1	1	1	1	1	1	1	0
3	2	1	1	1	1	1	1	1	0	0	0
4	3	1	1	1	1	1	1	1	0	0	0
5	4	1	1	1	1	1	1	1	1	1	0
6	5	1	1	1	1	1	1	1	0	0	0
7	6	1	1	1	1	1	1	1	0	0	0
8	7	1	1	1	1	1	1	1	0	0	0
9	8	1	1	1	1	1	1	1	0	0	0
10	9	1	1	1	1	1	1	1	1	1	0
11	10	1	1	1	1	1	1	1	0	0	0
12	11	1	1	1	1	1	1	1	0	0	0
13	12	1	1	1	1	1	1	1	1	1	0
14	13	1	1	1	1	1	1	1	0	0	1
15	14	1	1	1	1	1	1	1	0	0	0
16	15	1	1	1	1	1	1	1	1	1	0

Transpose

	A	B	C	D	E	F	G	H	I	J	K
1	Doc_ID	1	2	3	4	5	6	7	8	9	10
2	Doc_Relevancy	1	1	1	1	1	1	1	1	1	1
3	SQ01	1	1	1	1	1	1	1	1	1	1
4	SQ02	1	1	1	1	1	1	1	1	1	1
5	SQ03	1	1	1	1	1	1	1	1	1	1
6	SQ04	1	1	1	1	1	1	1	1	1	1
7	SQ05	1	1	1	1	1	1	1	1	1	1
8	SQ06	1	1	1	1	1	1	1	1	1	1
9	SQ07	1	0	0	1	0	0	0	0	1	0
10	SQ08	1	0	0	1	0	0	0	0	1	0
11	SQ09	0	0	0	0	0	0	0	0	0	0
12	SQ10	0	0	0	0	0	0	0	0	0	0
13	SQ11	0	0	0	0	0	0	0	0	0	0
14	SQ12	0	0	0	0	0	0	0	0	0	0
15	SQ13	0	0	0	0	0	0	0	0	0	0
16	SQ14	0	0	0	0	0	0	0	0	0	0

Fig. 2. Left picture is a sample of recording datasheet; Right picture is the transposed datasheet (input file)

Since it is generally expected that number of retrieved documents is much more than the number of search queries, it is suggested to assign columns to queries and rows to documents. However, the input file must follow the following criteria, which needs transposing the matrix of queries and documents (Fig. 2- the right picture):

- The file format must be in a comma separated values (CSV) format and with 'csv' extension.
- The columns represent the documents and the rows represent the search queries. Also, a header must be defined for each column and row.
- The first row represents the documents' IDs, which must be numerical and unique across the columns.
- The second row is designated for the relevancy of documents. The values must be numerical and can be either '0' or '1'.
- The third row and any row below that hold the search-queries match for each specific document. The values must be numerical and can be either '0' or '1'.
- The first column of the first two rows is the title and being ignored by the programme
- The first column from the third row onward represent the search-queries' IDs (Fig. 2- the right picture).

You can download a sample input file here.

4. Output File

The programme produces two types of outputs: log and statistical summaries. Firstly, there is only one log output file to save settings and runtime logs. Secondly, there are one or more statistical summaries in a comma-separated values (CSV) format. The CSV files hold the

optimized ranked list of search strategies based on the sensitivity (Fig. 3). The CSV files are suited to open in Microsoft Excel and record the combinations of search queries with their 'length', 'Sensitivity' and 'Precision'.

	A	B	C	D	E
1	#	Search Queries	Length	Sensitivity (%)	Precision (%)
2	1	SQ01_SQ03_SQ05_SQ07_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	12	90.049263	66.93518829
3	2	SQ01_SQ03_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	11	90.049263	66.93518829
4	3	SQ01_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	10	90.049263	66.93518829
5	4	SQ01_SQ05_SQ07_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	11	90.049263	66.93518829
6	5	SQ01_SQ03_SQ04_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	12	90.049263	66.91069031
7	6	SQ02_SQ04_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	11	90.049263	66.91069031
8	7	SQ01_SQ02_SQ04_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	12	90.049263	66.91069031
9	8	SQ01_SQ02_SQ03_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	12	90.049263	66.91069031
10	9	SQ01_SQ02_SQ03_SQ05_SQ07_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	13	90.049263	66.91069031

Fig. 3. A sample output file (when the minimize search strategy filter is off). The third row demonstrates the shortest highly sensitive search strategy, the maximizing version.

HSSS, sensitivity- and precision-maximizing version, can be formulated from the search queries presented in the second row or any row below that (when there is more than one combination of queries with the highest sensitivity) of a CSV file.

The best balance of sensitivity and precision, sensitivity- and precision-maximizing version of HSSS, can be obtained by finding the best-fit line on the scatter plot of sensitivity versus precision (see Fig. 2 in references 2 and 3). It is recommended to make the scatter plot from the output in the spreadsheet programme like MS Excel, while it has been filtered on "Report within Sensitivity Range of 100%" (Fig. 4).

5. Configuration

Before running the application, you must configure the application settings. The setting fields are as below:

Input File: The full path of the input file

Output Path: The full path of the directory that is going to be used for saving the outputs.

5.1. Data Configuration

5.1.1. Default: By choosing this option, SensPrecOptimizer calculates the sensitivity based on available number of documents and search queries in the input file.

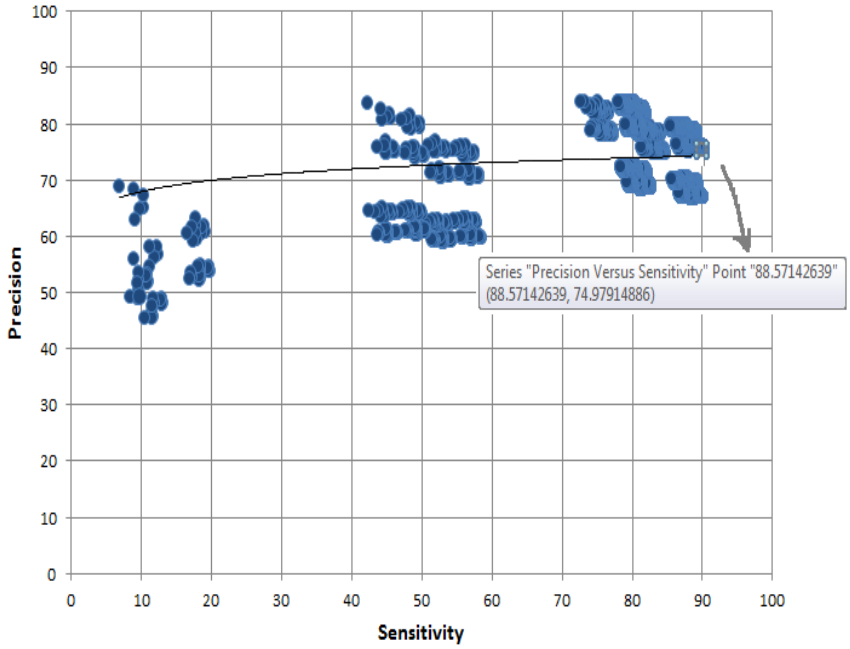


Fig. 4. Plot of sensitivity versus precision for different combinations of search queries from the sample output file; the intersection of trend line in the upper right hand of the plot and precision-sensitivity curve can be calculated as the optimization data points (as demonstrated for example).

5.1.2. Custom: Total number of relevant documents can be set when there is an external gold standard. For example, assume the input file consists of relevancy and retrieval status of 20 search queries within MEDLINE (4000 total documents and 2030 relevant documents) and there is a pool of known relevant documents retrieved by MEDLINE and EMBASE (4100 relevant documents). By choosing the DEFAULT option, the denominator of sensitivity formula (see the definition of sensitivity in Glossary) is 2030 and by customising this option, the denominator can manually be set as 4100.

5.2. Algorithm Configuration

The following options allow the user to modify the analysis process in a timely manner. It is recommended to set the algorithm configuration when the number of queries is more than 20.

5.2.1. Combinations Iterations - The Starting Point (%): The starting point is set to save the runtime and its memory usage, particularly when the number of search queries are larger than 20. It is recommended that the starting point set to a value below 40% or above 60%, because combination sizes close to 50% are memory and process intensive.

For example, if our input file has 20 search queries and we chose 75%, it means that SensPrecOptimizer initialised by 15 combinations.

5.2.2. Combinations Iterations - The Number of Additional Smaller Neighbours: The number of additional smaller combinations to be calculated after processing the smaller combinations (see Iteration 5 in the Technical Specification).

5.3. Runtime Filtering of Output

5.3.1. Default: The default minimum sensitivity sets to the maximum sensitivity of individual search queries.

5.3.2. Custom: The acceptable minimum sensitivity can manually be set to optimize the memory usage.

5.4. Configure Output Filter

The following options allow the user to customize the output.

5.4.1. Short Report by Minimizing Search Strategies' Length: It performs two types of filtering. Firstly, filter keeps the shortest search strategy within a set of queries combinations with similar sensitivity and precision. Secondly, it retains the strategy with the highest precision. For example, if combinations of 15, 16, 17 and 18 out of 20 queries have the same sensitivity and precision, choosing this option yields combinations of 15 queries only. It also filters the 15 queries combinations with the highest precision when there is different precision with the same sensitivity (Fig. 5).

	A	B	C	D	E
1	#	Search Queries	Length	Sensitivity (%)	Precision (%)
2	1	SQ01_SQ05_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	10	90.049263	66.93518829
3	2	SQ02_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	9	90	66.97213745
4	3	SQ01_SQ08_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	9	90	66.97213745
5	4	SQ01_SQ05_SQ08_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	9	89.950737	67.25598907
6	5	SQ01_SQ06_SQ07_SQ09_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	10	89.90148163	75.16474152
7	6	SQ01_SQ05_SQ07_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	9	89.75369263	75.6644516
8	7	SQ02_SQ07_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	8	89.70443726	75.71725464
9	8	SQ01_SQ07_SQ10_SQ11_SQ12_SQ13_SQ14_SQ15_	8	89.70443726	75.71725464
10	9	SQ01_SQ06_SQ07_SQ11_SQ12_SQ13_SQ14_SQ15_	8	89.65517426	75.92824554

Fig. 5. A sample output file filtered by minimizing search strategies' length

5.4.2. Report within Sensitivity Range of (%): The sensitivity range that is going to filter the outputs before outputting. It is recommended to set to a narrow range for the large search query sizes. It has been set by default to the range of 100%, which means a comprehensive report with no filter.

After you have set the configuration fields, you may click on the 'Run' button to execute the applications.

6. Error Messages and Troubleshooting

The following table summarizes the error messages and their diagnostics.

Error Message	Diagnostics
Memory Error	The programme is out of memory. Either it reached the program limit, the system does not have enough storage memory, or RAM left.
Algorithm is interrupted unexpectedly	An unexpected incident caused the algorithm to abort its process.
Invalid number of sub combinations	Invalid number of search queries in the input file
Invalid input file	The input file is invalid because it cannot be opened or it does not exist.
File structure could not be recognised	The input file must follow the CSV format and the supported language and Unicode are English and UTF-8. Please refer to the guideline for producing a valid input file.
Invalid file's parameter(s) defined	The followings might be out of range: • number of documents • number of search queries • total relevant documents
Invalid cell(s) in the file	An invalid column or row in the input file is detected
The output file cannot be accessed	The path to the output file is invalid. This issue might terminate the program without any warming (version 2.0.16 - Alpha).

References

[1] Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

[2] Mesgarpour B, Müller M, Herkner H. Search strategies-identified reports on ‘‘off-label’’ drug use in MEDLINE. J Clin Epidemiol 2012; 65: 827-834.

[3] Mesgarpour B, Müller M, Herkner H. Search strategies to identify reports on “off-label” drug use in EMBASE. BMC Med Res Methodol 2012; 12:190.

[4] Mesgarpour B, Mesgarpour M, Müller M, Herkner H. Developing software for combining search queries to design efficient search strategies. The Cochrane Library: Abstracts of the 19th Cochrane Colloquium 2011(supplement): 58.

Glossary

Boolean operators: Simple words used as conjunctions to combine or exclude keywords in a search, resulting in more focused and productive results. The most commonly Boolean operators used in search queries are “AND”, “OR” and “NOT”.

Controlled vocabulary: Consistent collection of terms chosen for specific purposes with explicit, logical constraints on intended meanings and relationships in a database. “Subject headings” make up the controlled vocabulary of a bibliographic database. For example, Medical Subject Headings (MeSH) in MEDLINE and EMTREE in EMBASE.

Free text: Words, phrases, or terms sought in title, abstract, or full text of document

Highly Sensitive Search Strategies (HSSS): The *Cochrane Handbook for Systematic Reviews of Interventions* offered two HSSS strategies for searching trials for inclusion in a Cochrane reviews: a sensitivity-maximizing version and a sensitivity- and precision-maximizing version [1].

Precision: The number of relevant identified reports divided by the total number of identified reports. Precision is equivalent to positive predictive value in diagnostic test method.

Proximity operators: They are used to improve the search queries by instructing the query to look for words that are within a short distance of each other in a document. Proximity operators and their functions vary by system and include “ADJ” and “NEAR”. Also called “adjacency operators”.

Retrieval performance: In this manual, used as a measure of sensitivity, precision or both.

Search filters: Filters are described as *predefined strategies* to improve recall and retrieve maximum recall of primary research of gold standard studies i.e. randomized trials (RCTs), systematic reviews, meta-analysis etc. and clinical queries such as diagnosis, prognosis, etiology and therapy. Filters are also referred to as hedges, PubMed Clinical Queries, or optimal search strategies.

Search query: One line in electronic search strategy. It is usually a text string containing the exact sequence of words and/or characters, and may include Boolean or proximity operators.

Search strategy: Several search queries connected with Boolean operators make up a search strategy.

Sensitivity: the number of relevant reports identified divided by the total number of relevant reports in existence. Sensitivity is also called “recall”.