Data Scaffolding: A Pragmatic Approach to Structuring a Literature Review Spreadsheet to Maximize Information, Facilitate Review, and Support Analysis

Kristine Ogden[1], Robert Hansen[1], Stacy Rattana[2], Nancy Neil[1].

[1] Lifecycle Sciences Group, ICON Clinical Research, San Francisco, CA
[2] Lifecycle Sciences Group, ICON Clinical Research, Chicago, IL

ICON
A Symbol of Excellence

Objective: Clinical literature reviews, especially systematic reviews, require the management and evaluation of large amounts of bibliographic data. Although the methodology for performing reviews is well documented, few sources share practical guidance on how to electronically manage citation data in order to efficiently inventory, organize, and assess published material. Spreadsheet tools (e.g., Excel®) give researchers unparalleled freedom to manipulate a large volume of information culled from source databases (e.g., MEDLINE®); however, many researchers lack a blueprint for structuring their data to facilitate review.

Methods: We describe a process of erecting "data scaffolding" that is critical to transforming a literature review spreadsheet into an agile, manipulable source of information that can triage, sort, and report on materials retrieved for review.

Results: Using a sample of literature data pulled from PubMed®, we describe several ideas for adding key structural components to a spreadsheet of citations that bring flexibility and functionality to the processes of article review and analysis. In doing so, we also describe the hows and whys of spreadsheet design to facilitate literature review, such as the importance of the spreadsheet layout; the addition of flag fields to use medical subject headings or key words to easily sort, organize, identify/retrieve, navigate, and summarize citations; and ways of leveraging built-in spreadsheet features such as autofilters, freeze-panes, and subtotals that can greatly facilitate the inventory and management of large volumes of citation data to support literature reviews.

Conclusions: In an era of increasing volumes of published literature and the proliferation of rigorous, formalized, systematic and non-systematic reviews, researchers must adapt to growing demands. Existing spreadsheet tools can be leveraged to effectively meet these demands, in ways that not only provide for greater efficiency but in ways that may result in enhanced research capabilities.

Introduction

Many researchers utilize spreadsheets to store and work with literature review data. Spreadsheets provide a versatile medium for interacting with large sets of complex data and offer a rich set of built-in features for performing data manipulation and analysis. However, choices in both spreadsheet data layout and structure (e.g., the amount of information in individual cells) greatly impact the number and types of strategies that may be leveraged to categorize, review, and analyze data.

We use the results of an illustrative MEDLINE® query to demonstrate data layout and other structural approaches that can help maximize the functionality of a Microsoft Excel® spreadsheet to support literature reviews.

Approach

A sample MEDLINE® search was developed and executed to identify quality of life and cost-effectiveness literature around the use of tamoxifen in the treatment of women with breast cancer, published on or before March 1, 2010.

PubMed® search of MEDLINE®:
("**breast neoplasms**"[MeSH Terms] AND "**tamoxifen**"[MeSH Terms]) AND ("**quality of life**"[Mesh] OR "**costs and cost analysis**"[MeSH Terms]) AND (("1"[PDat] : "**2010/03/01**"[PDat]))

The resulting PubMed® data (n=252 articles) were exported to an XML format and programmatically uploaded into an Excel® spreadsheet. Each data element (field) was loaded into a new column of the spreadsheet. Fields uploaded included: Year Published, Title, Journal (Full Name), Abstract, MeSH Keywords, MeSH Substances, Publication Type, Citation Subsets, PMID, PubMed® Status, Authors, Author Affiliation, and Language.

Spreadsheet Functions that Support the Literature Review Process

In the context of a literature review, a single spreadsheet can be used to support:

- **Initial review of abstracts**
  Spreadsheets can be used to collate resulting abstracts, track citations that do not meet the inclusion criteria, as well as the rationale for excluding a citation (i.e., exclusion coding). This ability to track and audit both included and excluded citations is particularly helpful in systematic reviews.
- **Full-text article inventory**
  Spreadsheets are vital when ordering, obtaining, and keeping track of full-text articles that are to be reviewed and included/excluded.
- **Reviewer assessment and consensus**
  This function is particularly helpful when multiple reviewers are required to conduct quality assessments of articles and/or data abstraction.
- **Data abstraction**
  The same spreadsheet can be used as a data abstraction tool. Data abstraction fields can be set up in the spreadsheet, next to the citation records, and populated with individual data points abstracted from each of the included citations.

Advanced Spreadsheet Features that Can Help Facilitate Review & Analysis

In addition to simply storing citation data, a spreadsheet can be set up to support the following functions:

- **Data filtering**
  Data can be hidden or shown, depending on the index terms assigned to a citation, the inclusion/exclusion status of a citation, a citation attribute (e.g., Year Published or Publication Type), or a data abstraction value. Multiple filters can be applied to a single spreadsheet, which is especially useful with large data sets containing several categories/types of data points.
- **Data sorting**
  Citation data can be organized and re-organized by key data fields (e.g., Year Published, Journal Name, Publication Type, etc.).
- **Data analysis**
  Once the data are abstracted, in many cases, the spreadsheet can be used to perform simple analyses (e.g., study and/or patient demographics, frequency counts of individual data points, etc.), as well as more complex analyses (e.g., statistical testing) of the abstracted data.
- **Decision logic**
  When information is organized into discrete cells, the data from multiple cells may be passed into a custom formula that runs decision logic. For example, if a user wanted to flag records that feature "quality of life" in the title or abstract and that were published in the last 5 years, it would be possible to write a custom formula that would identify each citation meeting these criteria.

Figure 1 and **Figure 2** show separate views of a sample spreadsheet of results from the sample breast cancer literature search described earlier. A filtered subset of the data is in view for clinical trial citations that are indexed for the MeSH term "Quality of Life".

In order to discuss the organization of the spreadsheet, section labels (A-E) have been inserted above groups of related data columns in Row 5.

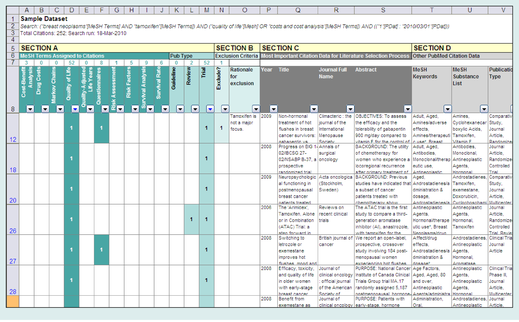

Figure 1: Screenshot of the Literature Spreadsheet with Scaffolding Fields

Description of Figure 1 Sections:

Section A – These are the data scaffolding fields added to the spreadsheet to flag citations that meet certain criteria (e.g., are indexed with a particular MeSH term, are indexed as a specific publication type, have been excluded during the literature review process, etc.). When a specific citation record (row) meets the flag field criteria, a "1" is entered in the cell so that all instances may be counted.

- The column headings (Row 8) have autofilters so that a user may hide rows that do not meet specific criteria or sort the dataset. In this example, the data have been filtered to show only those citations that are Clinical Trials (Column M=1) and indexed for the MeSH term "Quality of Life" (Column D=1).
- Subtotals (Row 7) located at the top of the dataset, recalculate whenever the sheet is filtered (calculated using the SUBTOTAL formula).
- As a visual cue, data scaffolding cells that display a "1" are shaded to so that it is easy for the user to tell how an individual citation is categorized (implemented using Conditional Formatting).

Section B – Two fields to record exclusion criteria; one field to code whether a citation is excluded and a second field to record the specific reason for exclusion. In column N, a "1" is recorded for each excluded citation so that all "1"s may be filtered to determine the total number of excluded citations.

Section C – These are the PubMed® data fields that are most helpful when applying inclusion/exclusion criteria during an initial review of the citations. To aid the reviewer, these fields are displayed together and placed near the flag field columns of indexed terms. Thus, the review doesn't have to scroll or jump across several columns to review related information.

Section D – These are PubMed® data fields that supplement the review of individual citations, but are not primary fields required for assessing initial inclusion/exclusion criteria. Fields listed in this section (not all are shown) include: MeSH terms, citation status, publication type, authors, author affiliation, language, etc.

Organizing the Literature Spreadsheet

In order to optimize the functionality of a spreadsheet and leverage several of the built-in functions that support data review and analysis, there are some basic guidelines for data layout that should always be considered **(Table 1)**.

Table 1: Basic Data Layout Guidelines for a Literature Review Spreadsheet

Guideline	Why?
Store all citation data in one spreadsheet	• Prevents mistakenly omitting observations because data are stored on separate sheets or in different external files • Improves efficiency reviewing and analyzing data because all the data are in one place • Summary statistics are obtained more readily if all literature data are housed in one place and calculations are limited to a single sheet • Allows easy access to excluded citations, which may often need to be re-evaluated
Orient data vertically (i.e., have column headings followed by rows of data stacked below the headings)	• Supports the use of built-in spreadsheet features for sorting and filtering data (e.g., using auto-filters to hide/show data rows based on specified criteria)
Organize the information for each unique citation in rows	• Prevents counting duplicate citations • Assists data filtering, ensuring only relevant, complete data are visible
Ensure that each column is a distinct field; arrange the columns in a logical sequence that facilitates review and analysis	• Segments the information into fields that may be used to sort & filter the spreadsheet (e.g., filter by Year Published, MeSH index terms, Publication Type, etc.) • Columns may be listed in an order that facilitates review and analysis tasks. For example, the fields for Year Published, Title, Journal and Abstract may be located together because they are all necessary to the initial citation review process • Possible field categories include: ○ Publication data ○ Inclusion/exclusion criteria ○ Quality assessment criteria ○ Data abstraction fields

Table 2 details the vertical organization of the spreadsheet data. By organizing groups of related data columns into vertical blocks (Sections A-E), these columns may be hidden if they are not relevant to a task. Subtotals are located at the top of the data set so that they are in always in view and are not accidentally sorted or mingled with individual citation records.

Table 2: Data Layout in Figures 1 and 2

Figure	Vertical Section of Figure	Spreadsheet Columns	Why?
1	**Section A** Flag Fields for Index Terms Assigned to Citations	A to M	• When reviewing abstracts, users may filter records by index term or publication type (i.e., all publications that are indexed to a particular subject may be reviewed and compared as a group) • Provides extra granularity to summary statistics about citations (e.g., it is possible to calculate the percentage of included/excluded articles indexed for a specific MeSH term).
1	**Section B** Internal Review Status Fields	N to O	• Allows a user to hide all rows that have been reviewed and/or excluded • Allows quick calculation of the number of excluded citations
1	**Section C** Primary PubMed® Fields (initial review)*	P to S	• Places information that is most important to the initial review in close proximity to the abstract being reviewed • Removes the need to jump around the spreadsheet or scroll across multiple columns.
1/2	**Section D** Other PubMed® fields*	T to AG	• The information is available for reference, but not cluttering the landscape.
2	**Section E** Abstraction Fields	AH to BA	• Places information for the analysis phase of the review in close proximity • Other blocks of unnecessary information may be hidden so that when abstraction is the task at hand, there is no unneeded information in view. • Improves the ease of creating an evidence table by being able to combine abstraction fields with PubMed® fields for bibliographic data in Sections C or D.

* Partially shown in Figures 1 and 2. Several data fields were collapsed to display the screenshot image.

Table 3: Features that May Be Leveraged to Categorize Spreadsheet Data and Improve Spreadsheet Navigation

Feature	Description	Purpose in a Literature Spreadsheet
autofilters	Autofilters are placed on a row of column headings and allow a user to sort a set of contiguous data, or hide/show data rows, based on specific criteria in one or more columns	• Enable sorting data by index term or Year Published. • Enable user to filter to show only included/excluded citations • Enable user to apply multiple criteria to filter a sheet (e.g., show all citations that are included, indexed for "Quality of Life" and published in 2009).
cell shading	Add background color to an individual cell	• Used in column headings as a visual cue for groups of columns that are related or useful to a particular task
conditional formatting	Add cell formatting based on criteria	• Used to apply shading to flag field column values that have a "1". Acts as a visual cue to show the flag criteria that apply to a particular citation.
FIND formula	Formula that searches a cell for a match to a letter or phrase. If the phrase is not found, the FIND formula returns an error message.	• Used in conjunction with an IF statement to build a formula to populate flag field columns by searching for whether an index term is associated with a particular citation (1=YES, 0=NO). • Can be used to search for the presence of specific words in a citation's Title or Abstract field, if needed.
freeze panes	Setting in which specific rows or columns remain visible when scrolling in the worksheet.	• Used to keep column headings and scaffolding columns in view when scrolling horizontally or vertically in the literature worksheet.
SUBTOTAL formula	Formula that can be set up to sort or count values in a spreadsheet range. Unlike the SUM formula, SUBTOTAL values will recalculate after autofilters are modified to display counts/sums of only the rows in view.	• Used to calculate total # s of citations that meet the criteria for a specific flag field (e.g., # citations indexed for "Quality of Life" or "Cost-Benefit Analysis").

Table 3 describes spreadsheet features that may be utilized to categorize and organize data. By using one or more of these features, not only can users improve spreadsheet navigation, but they can greatly assist in data management, review, and analysis. For example, by using autofilters, reviewers can quickly filter the literature results to those published in a specific year, or identify only those studies that enrolled a specific patient group or evaluated a specific clinical outcome (after data abstraction), or identify studies found to be of low quality (after quality assessment).

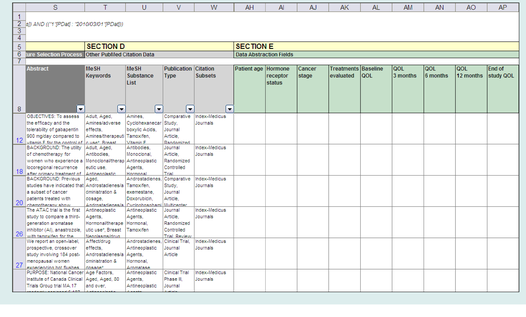

Figure 2: Screenshot of Data Abstraction Fields

Description of Figure 2 Sections:

Section D – These are PubMed® data fields that supplement the review of individual citations, but are not primary fields required for assessing initial inclusion and exclusion criteria. Fields listed in this section (not all are shown) include: MeSH terms, citation status, publication type, authors, author affiliation, language, etc.

Section E – This section includes columns for information that would be abstracted from citations that meet the inclusion criteria for the study. These data columns would change based on the particular aims and data collection requirements of the literature review.

CONCLUSIONS

There are a variety of strategies that may be employed to facilitate the inventory and management of large volumes of citation data in a spreadsheet for a literature reviews.

Past experience has shown that the incorporation of the scaffolding components outlined here has led to several key advantages/efficiencies:

- Efficiencies in the inventory, auditing, organization, and assessment of published material.
- The ability to easily sort, organize, and otherwise manipulate large volumes of information.
- The ability to address the growing demands for rigorous, formalized, systematic and non-systematic reviews.

Perhaps other researchers can take advantage of and build on these features, thereby making literature review spreadsheets a more integral, valuable, and empowering research tool.

ADDITIONAL RESOURCES

These references may be reviewed for more information about PubMed®, Microsoft Excel® and the organization of data in a spreadsheet format:

- Abdulezer L. Excel® Best Practices for Business. Indianapolis, IN: Wiley Publishing, Inc.; 2004.
- Microsoft Corporation. Excel® Home Page – Microsoft Office Online. Available at: http://office.microsoft.com/en-us/Excel/default.aspx. Accessed April 7, 2010.
- National Library of Medicine (NLM). PubMed® Online Training. Available at: www.nlm.nih.gov/bsd/disted/pubmed.html. Accessed April 7, 2010.
- Walenback J. Excel® 2003 Bible. Indianapolis, IN: Wiley Publishing, Inc.; 2003.