Estimating the Horizon of Articles to Decide When to Stop Searching in Systematic Reviews: An Example Using a Systematic Review of RCTs Evaluating Osteoporosis Clinical Decision Support Tools

Monika Kastner, PhD (cand)¹, Sharon Straus, MD, FRCPC¹,² Charlie H. Goldsmith, PhD³, ¹University of Toronto, Toronto, ON; ²University of Calgary, Calgary, AB; ³McMaster University, Hamilton, ON

Abstract
Researchers conducting systematic reviews need to search multiple bibliographic databases such as MEDLINE and EMBASE. However, researchers have no rational search stopping rule when looking for potentially-relevant articles. We empirically tested a stopping rule based on the concept of capture-mark-recapture (CMR), which was first pioneered in ecology. The principles of CMR can be adapted to systematic reviews and meta-analyses to estimate the Horizon of articles in the literature with its confidence interval. We retrospectively tested this Horizon Estimation using a systematic review of randomized controlled trials (RCTs) that evaluated clinical decision support tools for osteoporosis disease management. The Horizon Estimation was calculated based on 4 bibliographic databases that were included as the main data sources for the review in the following order: MEDLINE, EMBASE, CINAHL, and EBM Reviews. The systematic review captured 68% of known articles from the 4 data sources, which represented 592 articles that were estimated as missing from the Horizon.

Introduction
Although there are valid guidelines for improving the quality of reporting of meta-analyses (QUOROM)⁴, no guidelines exist on when to stop searching. One of the most important criteria in the QUOROM statement is to report on the data sources for searching for potentially-relevant articles, which include large bibliographic databases such as MEDLINE and EMBASE. However, there is currently no method that researchers can use to systematically and empirically decide when to stop searching when conducting systematic reviews.

Capture-mark-recapture (CMR) has commonly been applied to problems where there are multiple samples of some occurrence. However, not all of the population is observable in any one sample. Thus, the CMR process that can be applied involves getting an initial sample by some method of capture, marking the elements in the sample with some type of tag, and then putting the sample back into the population so that the marked elements are available to be recaptured in subsequent samples. This concept has commonly been used in ecology to estimate the number of fish in a lake²,³, where the first sample of fish is captured. The fish that are caught are then marked with a tag, and then released back into the lake and thus become available when a second and subsequent sample is taken from the lake. The number of marked fish in the second sample is then used with the sizes of the first two samples to estimate the number of fish in the lake. Although it was pioneered in ecology, the CMR concept is now being tested in epidemiology to check the completeness of case ascertainment in population-based studies⁴, and as a method for assessing publication bias in systematic reviews⁵. To our knowledge, one study has evaluated the completeness of systematic literature searches⁶. This study tested the CMR technique using the simplest model possible—the comparison of two databases (MEDLINE, and hand searching) for one journal. However, the process can be extended to more than 2 databases by using the confidence interval (CI) as a stopping rule of a search strategy, and reporting the estimate of the total number of articles and its CI as an estimate of the Horizon (i.e. all possible cases that could exist in the literature).

We tested a stopping rule retrospectively by applying the principles of CMR to estimate the Horizon of the total number of articles in the literature for a systematic review of RCTs, which evaluated clinical decision support tools for osteoporosis disease management.

Methods
Methods of the Systematic review:
Studies were identified by searching (in order) MEDLINE (1966 to July 2006), EMBASE (1980 to 2006), CINAHL (1982 to July 2006), and EBM Reviews (CDSR, ACP Journal Club, DARE, and CCTR). We also searched the grey literature: the websites of CIHR (Canadian Institutes of Health...
Research, AHRQ (Agency for Healthcare Research and Quality), CRISP (Computer Retrieval of Information on Scientific Projects), the US National Institutes of Health clinical trials register (ClinicalTrials.gov), BMJ Updates, Centre for Chronic Disease Prevention and Control, and Osteoporosis Canada. We reviewed the reference lists of relevant articles, hand searched the current clinical practice guidelines for the diagnosis and management of osteoporosis, and contacted experts in the field.

To generate our list of search terms, we conducted a preliminary search in MEDLINE and EMBASE using known terms and synonyms suggested by clinicians, librarians, and experts in the field to capture all possible textwords and MeSH terms that describe disease management and disease management tools. This list was supplemented by additional terms that were found in studies investigating disease management tools for other chronic diseases such as heart failure, diabetes, and asthma7-10. Using Ovid interface, these terms were first combined using the Boolean “OR”, and then “AND-ed” with osteoporosis in each database. The resulting retrieval yield was then limited using the best sensitivity search strategy filter for treatment studies developed by Haynes et al11.

Studies were included if they were RCTs in any language from 1966 to July 2006 that evaluated disease management interventions in men or women at risk for osteoporosis (age ≥ 65 years of age, postmenopausal women, or > 3 months systematic use of glucocorticoids), a confirmed diagnosis of osteoporosis, or with an existing or previous fragility fracture. Interventions could be in any format (e.g. electronic-, paper-, or program-based) as long as they incorporated an aspect of care coordination and were characterized by any one or a combination of terms that represent disease management tools. The intervention could be directed to patients or any health care professionals in the continuum of care of patients; had to evaluate live patients; and included any component of osteoporosis disease management (i.e. risk assessment, diagnosis, or treatment). We excluded any interventions that evaluated pharmacological strategies (e.g. bisphosphonates, hormone replacement therapy, selective estrogen receptor modulators, or vitamins), unless they were a component of the disease management tool. Outcomes of interest were fragility fractures (vertebral or nonvertebral), bone mineral density (BMD) investigations, initiation of any osteoporosis treatment, and fracture-related complications (e.g. quality of life, admission to long-term care, and fracture-related mortality). We excluded any outcomes related to fractures due to major trauma, primary prevention of osteoporosis, and falls prevention.

Two investigators (MK and SS) independently reviewed the titles and abstracts of potentially relevant citations, applied the inclusion/exclusion criteria, and rated each article as “relevant”, “not relevant”, or “unsure” using a standardized form. To identify the final list of studies to be included in the systematic review, the same 2 investigators independently applied the inclusion/exclusion criteria for articles that were selected for full-text review. Using a standardized data abstraction form, 2 reviewers (MK and SS) independently extracted data on the setting, study design (method of randomization, allocation concealment, and blinding); population characteristics (inclusion/exclusion criteria, sample size, number of patients assessed for eligibility and the number who met inclusion criteria); interventions; outcomes; results; and follow-up (duration of follow-up, number of patients followed-up, intention to treat (ITT) analysis, number of participants withdrawn or lost to follow-up, and if any differences existed between groups for drop-out rates).

Figure 1: Study identification Flowchart
We assessed study quality using specific methodological criteria that are considered the most relevant to the control of bias: randomization, random allocation concealment, blinding, and completeness of follow-up (including ITT analysis, withdrawals, and reasons for dropouts).

Our search of MEDLINE, EMBASE, CINAHL, and EBM Reviews identified 39,953 potentially relevant citations—none were identified from the grey literature. Once duplicates and non-randomized studies were removed, MK and SS independently screened the titles and abstracts of 1246 RCTs for relevance. Of the 42 articles that were selected for full-text review, 14 RCTs met the inclusion/exclusion criteria and were included in the systematic review.

**Methods of Horizon Estimation**

The 1246 potentially relevant articles were organized in an Excel database to document whether each of these articles were captured by the 4 databases (MEDLINE, EMBASE, CINAHL, or EBM Reviews) by indicating “Y” for yes, and “N” for no. The data were transferred to Egret®¹², which is a statistical program that can calculate the number of missing articles and its 95% CI with a Poisson regression model.

To estimate the population of articles of unknown size N, the basic principle behind the CMR concept is that there are k independent samples of database searches of size n, where i=1,2,…,k. The task is to estimate N, which is the horizon of articles. After each search, the known relevant articles are marked since they can be put into a literature database. They are available for identification by another database search because they have not been removed from the population of known references (i.e., they do not need to be “put” back into the pool of references).

If Mi is the number of distinct articles known at the end of the i-th search while mi is the number of new articles detected by the i-th, then M_i=m_i, M_2=m_1 +m_2, or M_k=sum of the m_i as i=1 to k. At the end of the k-th search, 2^k–1 cells for the search are known and 1 cell is unknown. The unknown cell is the count of the number of articles in the horizon that are not detected by at least one of the k data base searches.

If Poisson regression models permit the estimation of the missing cells when a model is fitted to the known cells, the modeling strategy can be used to estimate the number of articles missed by the current searches, say, a_i. The modeled cells permit the estimation of a_i, and once a_i is estimated then N can be estimated by adding the estimate of a_i to M_i (i.e., Est (N_i)=M_i+a_i, i ≥ 2). Poisson regression can be used to estimate a_i as well as to provide estimates of the variation to create CIs.

Egret provides estimates of a_i and its 95 % CI using a proportional fitting algorithm that is iterative. Model fit can be assessed by comparing the estimates of the fitted cells with the observed cells. Once the best model is selected, the missing cell estimate is used as the estimate of the number of missed articles. This estimate plus the M_i known articles are added to give the horizon estimate. Both the missing cell and the horizon estimate can be bounded with 95% CIs using the Egret output directly.

**Results**

The results of the horizon estimation for the systematic review are shown in Figure 2. The horizon was estimated to be 1729 articles (95% CI 1636 to 1839) after searching the first two databases (MEDLINE plus EMBASE), and 1739 articles (95% CI 1621 to 1846) after searching the first 3 databases (MEDLINE plus EMBASE plus CINAHL). The results showed that the horizon estimation after searching all 4 databases was 1838 articles (95% CI 1749 to 1955), which represented an estimated 1838–1246=592 articles that were missing from the known number of articles.

**Discussion**

The result of this horizon estimate indicates that the systematic review captured 68% of known articles from 4 large bibliographic databases. This result has several implications, and has to be interpreted with caution for several reasons.

First, the horizon estimation procedure has never been tested on a systematic review with ≥ 2 data sources.
sources, so it is not possible to compare it with any other results. It is possible that our findings represent as good or better results than would be expected for other rigorously conducted systematic reviews. More studies are needed to evaluate the process so that we can derive and validate a definition to guide the interpretation of the horizon estimate in terms of how close we need to get to the Horizon to be credible.

Second, we tested the horizon estimation retrospectively (i.e., there was no empirical rationale for the order of searching we selected: MEDLINE, EMBASE, CINAHL, and EBM reviews), which did not enable the analysis to be iterative. One of the advantages of the horizon estimation (if performed prospectively) is that the sequence of databases searched can inform a more optimal order of searching and the decision to either continue searching subsequent databases or to stop if the known number of articles falls within the confidence intervals of the horizon estimate. Our calculations indicated that the optimal search order would have been MEDLINE, EMBASE, EBM reviews, and CINAHL because it arrived at the same 1246 articles with 4 more articles known at one search earlier than they would have been with the original order. However, it is unlikely that a prospective evaluation of our search strategy would have changed our decision to discontinue searching all of the databases since our known number of articles did not fall within the horizon estimate CI during any of our hypothetical steps. Furthermore, all of our known studies were identified from the first 4 databases tested in this report, even though our search extended beyond these sources (i.e., the grey literature, hand searches of relevant journals, and contact with experts in the field). This approach would typically be considered an exhaustive search, which raises some questions about which other sources we might have searched to find the 592 articles that were identified as missing from the Horizon, or if the missing articles represent unpublished studies.

Third, the missing articles may represent studies we excluded at first screening for relevance. This raises an important question to consider—what is the most appropriate point during the article selection process to calculate the horizon estimation, to ensure that we get an accurate representation of the Horizon? Our estimation was based on data that are representative of 1246 abstracts that were screened independently for relevance by 2 investigators. It might be possible that the estimate would have been closer to the Horizon if the calculation were performed on the set of unscreened articles that were captured from one step earlier in the article selection process (i.e. the 3880 potentially relevant RCTs, see Figure 1). Our future work will involve testing this hypothesis.

Lastly, the systematic review that was used as an example in this report was done rigorously. It incorporated all the quality criteria for systematic reviews as outlined by the QUOROM statement, and used traditional methods for completing the review. We conducted a comprehensive literature search with well-defined terms and inclusion/exclusion criteria, selected articles independently at each level of article, and addressed potential sources of variability between relevant studies (in addition to random error) by the rigorous assessment and identification of the methodologic quality of studies (randomization, blinding, and follow-up); and differences of population, interventions, and outcomes. The 1246 potentially relevant articles represent 99% of the initial search yield of nearly 40,000 articles that was found by our search strategy. This raises questions about whether traditional searching methods are sufficient or if we need to change our searching strategies. If so, how can we improve our searching, and at what point would we be willing to say that our search is sufficient?

Other applications of CMR
The CMR concept can be applied to establish a stopping rule for searching in systematic reviews. It has been applied in epidemiological studies to estimate the number of cases of disease such as myocardial infarction and inflammatory bowel disease. Other recent applications of the horizon estimation include the use of this concept to calculate the horizon of journals that were missed to help define a journal subset for nephrology content.

Conclusion
The CMR technique can estimate the number of missing articles and the horizon of articles in a systematic review or meta-analysis by their 95% CIs. However, more studies are needed to objectively define cutpoints for these estimates that can be used to interpret how close the search can approach the Horizon; and to evaluate the effectiveness of this estimation as a stopping rule for searching.

References
1. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup F, for the QUOROM Group. Improving the quality of reports of meta-analyses...


