Guidelines for Systematic Review in Conservation and Environmental Management

ANDREW S. PULLIN* AND GAVIN B. STEWART

Centre for Evidence-Based Conservation, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

Abstract: *An increasing number of applied disciplines are utilizing evidence-based frameworks to review and disseminate the effectiveness of management and policy interventions. The rationale is that increased accessibility of the best available evidence will provide a more efficient and less biased platform for decision making. We argue that there are significant benefits for conservation in using such a framework, but the scientific community needs to undertake and disseminate more systematic reviews before the full benefit can be realized. We devised a set of guidelines for undertaking formalized systematic review, based on a health services model. The guideline stages include planning and conducting a review, including protocol formation, search strategy, data inclusion, data extraction, and analysis. Review dissemination is addressed in terms of current developments and future plans for a Web-based open-access library. By the use of case studies we highlight critical modifications to guidelines for protocol formulation, data-quality assessment, data extraction, and data synthesis for conservation and environmental management. Ecological data presented significant but soluble challenges for the systematic review process, particularly in terms of the quantity, accessibility, and diverse quality of available data. In the field of conservation and environmental management there needs to be further engagement of scientists and practitioners to develop and take ownership of an evidence-based framework.*

Keywords: conservation policy, conservation practice, decision making, evidence-based knowledge transfer

Directrices para la Revisión Sistemática en Gestión Ambiental y de Conservación

Resumen: *Un mayor número de disciplinas está utilizando marcos de referencia basados en evidencias para revisar y diseminar la efectividad de las intervenciones de gestión y política. El fundamento es que la mayor accesibilidad de la evidencia mejor disponible proporcionará una plataforma de toma de decisiones menos sesgada y más eficiente. Argumentamos que hay beneficios significativos para la conservación al utilizar tal marco de referencia, pero la comunidad científica debe emprender y diseminar revisiones más sistemáticas antes de que se pueda comprender el beneficio completo. Diseñamos un conjunto de directrices para realizar revisiones sistemáticas formales, basado en un modelo de servicios de salud. Las etapas de las directrices incluyen la planificación y conducción de una revisión, incluyendo formación del protocolo, estrategias de búsqueda, inclusión de datos, extracción y análisis de datos. La diseminación de revisiones es abordada en términos del desarrollo actual y los planes futuros para una biblioteca de acceso abierto en la Web. Al utilizar estudios de caso resaltamos modificaciones críticas a las directrices para la formulación del protocolo, evaluación de la calidad de los datos, extracción de datos y síntesis de datos para la gestión ambiental y de conservación. Los datos ecológicos presentaron retos significativos, pero solucionables, para el proceso de revisión sistemática, particularmente en términos de la cantidad, accesibilidad y calidad de los datos disponibles. Se requiere un mayor compromiso de científicos y profesionales de la gestión ambiental y de conservación para desarrollar y apropiarse de un marco de referencia basado en evidencias.*

Palabras Clave: política de la conservación, práctica de la conservación, toma de decisiones, transferencia de conocimiento basado en evidencia

*email a.s.pullin@bham.ac.uk
Paper submitted August 4, 2005; revised manuscript accepted January 25, 2006.

Conservation Biology Volume 20, No. 6, 1647–1656

DOI: 10.1111/j.1523-1739.2006.00485.x

Introduction

In response to problems of accessing scientific information to support decision making, many applied disciplines are utilizing an evidence-based framework for knowledge transfer involving systematic review and dissemination of evidence on effectiveness of interventions at the practical and policy levels (Stevens & Milne 1997; Khan et al. 2003). The framework is most fully developed in the health services sector, where global review and dissemination units have been established and are linked by networks such as the Cochrane Collaboration (e.g., www.cochrane.org). Within these networks systematic reviews are undertaken following set guidelines that include peer review to ensure that they meet required standards before dissemination. The need for such a framework in conservation has been argued elsewhere (Pullin & Knight 2001; Fazey et al. 2004; Pullin et al. 2004; Sutherland et al. 2004). Here we present a summary of newly developed guidelines for systematic review and dissemination in conservation and environmental management (more detailed guidance can be obtained from www.cebc.bham.ac.uk).

We used established guidelines from the health services sector (NHS CRD 2001; Higgins & Green 2005) as our models, undertook our own systematic reviews to test these models, and modified the guidelines through analysis of procedures and outcomes for their application to conservation and environmental management. Although the basic ethos of systematic review remains unchanged, ecological data are often fundamentally different in nature from data on human health (Fazey et al. 2004; Pullin et al. 2004), and this is reflected in our guidelines. At first glance many of the guidelines may seem routine and common sense, but the rigor and objectivity applied at key stages, and the underlying philosophy of transparency and independence, sets them apart from the majority of traditional reviews published recently in the field of applied ecology (Roberts et al. 2006). Pullin and Knight (2001), Fazey et al. (2004), Pullin et al. (2004), and Sutherland et al. (2004) argue that, once established, systematic review methodology will significantly improve the identification and provision of evidence to support practice and policy in conservation and environmental management. For this methodology to have an impact on conservation effectiveness, more conservation biologists need to undertake reviews, and we encourage this community to use (and improve) these guidelines and help establish an evidence-based framework for our discipline.

Systematic Review Guidelines

For clarity the guidelines are split into three stages and key phases within each. We use examples of our own reviews to highlight key issues for reviews in conservation and environmental management.

Stage 1—Planning the Review

QUESTION FORMULATION

A systematic review starts with a specific question, clearly defined with subject, intervention, and outcome elements (Table 1), that is answerable in scientific terms (Jackson 1980; Cooper 1984; Hedges 1994). The question is critical to the process because it generates the literature search terms and determines relevance criteria (NHS CRD 2001). Finding the right question is a compromise (probably more so in ecology than in medicine) between taking a holistic approach (thus increasing realism by involving a large number of variables but limiting the number of relevant studies), and a reductionist approach (which may limit the review's relevance, utility, and value) (Stewart et al. 2005*a*). The question should be practice or policy relevant and should therefore be generated by, or at least in collaboration with, relevant decision makers (or organizations) for whom the question is real. It may also be important for the question to be seen as neutral to stakeholder groups. Ideally meetings should be held with key stakeholders to try to reach consensus on the nature of the question. This may be more critical for ecological review than medical review because, unlike the benefit of improving human health, the benefit of conserving biodiversity is often contested (Fazey et al. 2004).

Table 1. Elements of a reviewable question; normally a permutation of "Does intervention x on subject y produce outcome z"?

Question element	*Definition*
Subject	unit of study (e.g., ecosystem, habitat, species) that should be defined in terms of the subject(s) on whom the intervention will be applied
Intervention	proposed management regime, policy, or action
Outcome	all relevant objectives of the proposed management intervention that can be measured reliably with particular consideration given to the most important management outcome and to any outcome critical to whether the proposed intervention has greater benefits or disadvantages than any other alternatives (i.e., the outcome desired)
Comparator	Is the intervention being compared with no intervention or are alternative interventions being compared with each other?

EXAMPLES OF QUESTION FORMULATION

We use four examples of the systematic review process throughout and introduce each of them here.

Example 1. English Nature, a U.K. statutory conservation agency, was concerned about the ecological impacts of burning management carried out by landowners in upland areas of England. Discussion with English Nature personnel enabled this general concern to be "unpacked," allowing definition of subject, intervention, and outcome elements of two specific review questions (Stewart et al. 2005*a*): "Does burning of U.K. submontane, dry dwarf-shrub heath maintain vegetation diversity?" and "Does burning degrade blanket bog?" Identification of these two related questions allowed specific hypotheses to be tested while retaining broader policy relevance. These also provided examples of habitat-based reviews.

Example 2. The Royal Society for the Protection of Birds (RSPB) was concerned about the impact of wind farms on bird populations, which led to a systematic review (Stewart et al. 2005*b*). This review was a test case for measuring impact of interventions arising from specific development activity with policy relevance.

Example 3. Tyler and Pullin (2005) and Tyler et al. (2006) examined the effectiveness of *Rhododendron ponticum* control methods as a test case for review of control methods for an invasive plant species.

Example 4. Tyler et al. (2005) investigated the impact of control methodologies on introduced populations of the American mink (*Mustela vison*) in Europe, as an invasive animal species test case.

Although discussions with the proposers of a review proved effective in formulation of a review question, other stakeholders may disagree. In example 1, a key stakeholder disagreed with the outcome measure (a measure of favorable ecological condition based on the relative abundance of key species) used in the "blanket bog" review. To avoid postreview problems such as this we advocate involvement of multiple stakeholders early in the review process.

DEVELOPING A REVIEW PROTOCOL

The review protocol acts as a document that all stakeholders agree upon, after which the review itself can be conducted (see www.cebc.bham.ac.uk/protocols.htm for examples).

A review protocol is developed as a document that guides the review. As in any scientific endeavor, methodology should be established and made available for scrutiny and comment at an early stage. Because reviews are retrospective by nature, the protocol is essential to make the review process as rigorous, transparent, and well defined as possible (Light & Pillemer 1984). Besides a formal presentation of the question and its background, a review protocol sets out the strategy for obtaining data and defines relevance criteria for data inclusion or exclusion (NHSCRD 2001). The subject, intervention, and outcome elements defined in the question-setting stage provide a priori inclusion criteria. If the relevant population, intervention, or outcome measures are present, then the data are included although data quality thresholds may result in the subsequent exclusion of otherwise relevant material either from quantitative analysis or from the review in entirety (see below).

The search strategy is constructed from search terms extracted from the subject, intervention, and outcome elements of the question. It is important that the search is sufficiently rigorous and broad so that all studies eligible for inclusion are identified. Search protocols must balance sensitivity (getting all information of relevance) and specificity (the proportion of hits that are relevant) (NHS CRD 2001). In ecology resource-intensive searches of high sensitivity are required, even though this is at the expense of specificity, because ecology lacks the mesh-heading indexes and integrated databases of medicine and public health. A high-sensitivity and low-specificity approach is necessary to reduce bias and increase repeatability (see below). Typically, large numbers of references are therefore rejected. For example, of 317 articles with relevant titles concerning the impact of burning on blanket bog, only 8 (2.5%) had comparators (Stewart et al. 2005). Similarly, reviews regarding burning of dry heath and the impact of wind farms on bird abundance resulted in meta-analysis of 1.7% and 12% of material with relevant titles, respectively.

In a review of the effectiveness of control methodologies on introduced populations of the American Mink (*Mustela vison*) in Europe, Tyler et al. (2005) searched the following electronic databases: Agricola, BIOSIS previews, CAB abstracts, Copac, Digital Dissertations, Index to Theses online, ISI Current Contents, ISI Proceedings, ISI Web of Knowledge, ISI Web of Science, JSTOR, ScienceDirect, Scirus, Scopus, Wildlink; the World Wide Web (first 100 "hits" from www.alltheweb.com, www.google.co.uk, U.K. Department for the Environment, Food and Rural Affairs, Scottish Natural Heritage, Oxford University's Wildlife Conservation Research Unit, The Royal Society for the Protection of Birds, The National Trust, British Wildlife, The Mammal Society, Mammals Trust, and The British Trust for Ornithology); and bibliographies of relevant articles (search terms: *Mustela* AND *vison*, *Mustela* AND *vison* AND trap*, *Mustela* AND *vison* AND control*, *Mustela* AND *vison* AND management, *Mustela* AND *vison* AND pest, Mink AND trap*, Mink AND control*, Mink AND management, Mink AND pest). The specificity of this search was low, with many references identified multiple times. The grey literature search was largely U.K. based due to resource limitations, although the inclusion of non-U.K. theses was possible. The low specificity of the review (only 1% of retrieved material was judged relevant), however, limits the potential for bias notwithstanding the geographical scope of the grey-literature

search. The documented search is fully repeatable and transparent; thus, readers can judge its validity.

Stage 2 - Conducting the Review

SEARCHING FOR DATA

It is perhaps self-evident that the widest possible range of sources should be accessed to capture information. The following are useful general sources: multiple electronic databases (general databases and databases with specific foci), professional networks and organizations (special-interest groups may have personal literature collections or libraries), the Internet (method of contacting or searching for information from the above two groups), bibliographies (data sources cited in literature obtained from the above), and subject experts (direct personal contact may yield new data sets). To minimize the problem of publication bias (e.g., Leimu & Koricheva 2005), both published and unpublished data must be included, a standard rarely satisfied in traditional reviews. Hand searching of specific sources and visits to libraries and museums are likely to be necessary to extract all relevant material. It may be necessary to search local databases for questions with a regional focus. At each stage of the review it is essential that the numbers and identities of articles retrieved, accepted, and rejected be recorded. The maintenance of a database or collection of bibliographic software libraries is recommended. The repeatability of search methods is a key characteristic of systematic reviews (NHS CRD 2001).

SELECTION OF RELEVANT DATA

Once searching is complete, relevant articles must be efficiently selected without wasting resources examining irrelevant articles in detail. Selecting only relevant articles from a potentially large body of initial literature requires the reviewer to use inclusion and exclusion criteria stated a priori in the protocol to impose a number of filters of increasing rigor. First, if a long list of articles or data sources is acquired (1000s rather than 100s) and the list of relevant sources is likely to be much shorter, it may be efficient to eliminate some material on title only (especially if obviously spurious hits arise from ambiguity in the use of words in the literature). The second filter should examine title and abstract to determine relevance. The approach should be conservative so as to retain data if there is reasonable doubt over its relevance. It is good practice at this stage to employ a second reviewer to go through the same process on a random subsample of abstracts from the original list and to ensure decisions are comparable by performing a kappa analysis, which adjusts the proportion of records for which there was agreement by the amount of agreement expected by chance alone (Cohen 1960; Edwards et al. 2002). If comparability is not achieved, then the criteria should be further developed and the process repeated.

Remaining articles should be viewed in full to determine whether they contain relevant and usable data. Obtaining the full text of all articles can be very time consuming and a realistic deadline may have to be used and a record kept of those not obtained. The conservative approach and independent checking of a subsample by kappa analysis should be repeated at this stage. Short lists of articles and data sets should be made available for scrutiny by stakeholders and subject experts. All should be invited, within a set deadline, to identify relevant data sources they believe are missing from the list. Reviewers should be aware that investigators often cite selectively studies with positive results (Gotzsche 1987; Ravnskov 1992); thus, checking bibliographies and direct contacts must be used only to augment the search.

ASSESSING QUALITY OF METHODOLOGY

To determine the level of confidence that may be placed in selected data sets, each should be critically appraised to determine the extent to which its research methodology is likely to prevent systematic errors or bias (Moher et al. 1995). In the health services, a hierarchy of research is recognized that scores the value of the data in terms of the scientific rigor of the methodology used (Stevens & Milne 1997). The hierarchy of methodology can be viewed as generic and has been transferred from medicine to ecology (Pullin & Knight 2003; see www.cebc.bham.ac.uk for full details). Where a number of well-designed, high-quality studies are available, others with inferior methodology may be rejected. Alternatively, the effects of individual studies can be weighted according to their position in the "quality hierarchy." However, there are dangers in the rigid application of this hierarchy in ecology. Hypothetically, a rigorous methodology, such as a randomized controlled trial, could be viewed as superior, even though it was applied over inadequately short time and small spatial scales, to a time series experiment providing data over longer time and larger spatial scales more appropriate to the question. This problem carries with it the threat of misinterpretation of evidence. Potential pitfalls of this kind need to be considered at this stage and addressed by more pragmatic quality weightings and judicious use of sensitivity analysis (see below).

Four sources of systematic bias are routinely considered in healthcare (Feinstein & Horwitz 1985; Moher et al. 1995; Moher et al. 1996; Khan et al. 2003) of which three have, to date, required consideration in ecological systematic reviews. Selection bias results from the way that comparison (e.g., treatment and control) groups are assembled (Kunz & Oxman 1998) and is a primary reason for randomization. Performance bias refers to systematic differences in the care provided to subjects in the comparison groups and is dealt with by the experimenter

being unaware of which are treatments and which are controls (blinding) (Schulz et al. 1995). We postulate that the ecological equivalents of performance bias arise from biased baseline comparisons and failure to consider the impact of covariables along with the intervention of interest. However, it is not possible to account for variables that are not known to be confounders or that were not measured, and for those that are known, difficulties can arise in extracting standardized information for analysis. Measurement or detection bias refers to systematic differences between the comparison groups in outcome assessment and is also addressed by blinding (Schulz et al. 1995). Blinding is generally not possible in ecology, but detection bias nevertheless varies, depending on the rigor and objectivity of sampling methodology (e.g., percent cover assessed by eye is subject to greater potential detection bias than frequency). The fourth, attrition bias (systematic differences between the comparison groups in the loss of samples), has not been an issue in ecological systematic review to date.

Assessing the quality of methodology is a critical part of the systematic review process and requires a number of subjective decisions about the relative importance of different sources of bias and data quality elements specific to ecology, particularly the appropriateness of variable temporal and spatial scales. It is therefore vital that the assessment process be standardized and be as transparent and repeatable as possible. At least 25 scales and 9 checklists have been used to assess the validity of randomized controlled trials in medicine (Moher & Feinstein 1995; Moher et al. 1996), and various similar criteria have been used to critically appraise the validity of observational studies (Horwitz et al. 1979; Feinstein et al. 1982; Levine 1994; Bero et al. 1999). These checklists do not consider specific ecological criteria. We therefore suggest that review-specific a priori assessment forms and two or more assessors should be used to assess study quality in ecological reviewing. The subjective decisions may be a focus of criticism; thus, we advocate consultation with stakeholders to try and reach consensus before moving on to data extraction.

Finally, at this stage it may be necessary to reject articles that are seemingly relevant but do not present data in extractable format. If possible, authors of such articles should be contacted and asked whether they can provide data in a suitable format.

Stewart et al. (2005*a*) used the hierarchy of methodology to separate randomized controlled trials and site comparisons addressing the question, "Does burning degrade blanket bog?" This reflected a major data-quality schism; therefore, further data-quality assessment was inappropriate given the very small number of studies. This approach enabled a simple, but discriminatory, vote count of studies with results showing positive, neutral, or negative effects.

When reviewing the impact of wind farms on bird populations, the standard hierarchy of evidence was considered inadequate by itself due to variation in other critical data-quality elements, particularly the widespread occurrence of confounding factors resulting from variation between treatment and control at baseline or from changes concurrent with wind-farm operation (ecological performance bias). The rigor of observations was also variable as measured in terms of replication and objectivity (ecological detection bias). To test for the impact of these factors, data-quality scores, summing the different aspects of data quality outlined above, were added as a meta-regression covariable. Data-quality score was not significant, suggesting that bifurcation of the data into high- and low-quality evidence was not necessary, possibly because the low-quality studies (low replication, imprecise estimates of abundance, high intratreatment variation coupled with confounded baselines) had a high variance and therefore a low weighting in meta-analysis by inverse variance. Sensitivity analyses were used to explore the impact of including low-quality unreplicated data, but the impact of individual data quality elements other than time was not examined because a large number of environmental and wind-farm correlates were of interest and the potential for Type II errors would have been increased. Although this pragmatic approach is easy to apply, there is no measure of a study's "true" validity (Emerson et al. 1990; Schulz et al. 1995; Jüni et al. 1999). Caution should be exercised in interpreting study validity, especially if different quality elements are combined in a single data-quality sum.

A review of the effectiveness of *Rhododendron* control methods considered study hierarchy and potential for bias providing a subjective summary of data quality (Table 2). In this instance the number of environmental variables with sufficient data for analysis was low and sample sizes were sufficient to examine the impact of some individual study quality variables such as length of experiment and whether results were generated in the field or a glass house. There were statistically significant differences in effectiveness of control with glasshouse trials showing greater control than field-based experimentation or monitoring, raising questions about the ecological relevance of glasshouse work and the likely modifying variables. This approach has the merit of objectivity, although there is choice about which variables are included in the analysis and caution must be exercised to avoid Type II errors, data mining, and overinterpreting results, especially when sample sizes are small.

DATA EXTRACTION

Data extracted from articles should be recorded on carefully designed spreadsheets and undertaken with synthesis in mind. Narrative synthesis requires the construction of tables that provide details of the study or population characteristics, data quality, and relevant outcomes, all of which are defined a priori. Quantitative analysis follows the same model but care must be taken to extract

Table 2. Data-quality assessment of an article included in a systematic review of the effectiveness of methods for the control of *Rhododendron ponticum* (Tyler et al. 2004).

Methods	site comparison based on sites treated with different interventions, no control, comparison methods only
Population	no stand-age detail, site located on lowland heath
Intervention and cointerventions	drilled holes filled with herbicide, compared with stumps painted with herbicide
Outcomes	painted stumps, 30–40% killed drilled holes, 95% killed
Study design	site comparison
Baseline comparison	no information regarding the sites prior to treatment, thus not possible to validate baseline
Intratreatment variation	no information describing intratreatment variation
Measurement of intervention and cointerventions	no information regarding the sites provided, thus not possible to comment on other management within the area
Replication and parameter of abundance	no replication or measure of abundance other than percent kill
Notes	study appears to comment on the use of techniques rather than providing the reader with scientific evidence, resulting in a high potential for bias and subsequently low data quality

information pertinent to subsequent analysis (e.g., should binary or continuous outcomes be extracted)? In contrast to medicine, consideration of the appropriate spatial scale(s) and level of replication are necessary prior to extracting the variance measures required to weight meta-analyses. Great care must be taken to standardize and document the process of data extraction, the details of which should be recorded in tables of included studies to increase the transparency of the process. To some extent data extraction can be guided by a priori rules, but the complexity of the operation means a degree of flexibility must be maintained. Sensitivity analyses can be used to investigate the impact of extracting data in different ways when there is doubt about the optimum extraction method.

Reviewing the impact of burning on the ecological condition of blanket bog required extraction of data showing changes in floristic composition and structure. Two reviewers extracted data after reaching a consensus regarding which subsets were relevant within the full data set of each article. A priori rules increased the repeatability of data-set formation. For example, sites within an experiment were pooled to prevent pseudoreplication, avoiding post hoc justifications for deriving more than one data set from an experiment and combining unreplicated, pseudoreplicated, and replicated data. Pooled treatment and control sites were included once to maintain independence and avoid bias, with the exception of data on rotational burning, which was scarce and therefore admitted to the review provided there was a comparator irrespective of further potential for bias. Where there was a choice of times since burning, priority was given to the longest time range to maintain independence and maximize predictive power. Similarly, grazed sites received priority over ungrazed sites when the maintenance of independence demanded a choice because grazing and burning are carried out concurrently over most of the British uplands (Stewart et al. 2005*a*). If sample sizes had been larger and a quantitative generic outcome measure identified, the impact of these decisions could have been explored with sensitivity analyses. Given the nature of the data, qualitative discussion of the issues was more appropriate.

DATA SYNTHESIS

This stage includes both qualitative synthesis and quantitative analysis with statistical methods as appropriate. Qualitative synthesis allows informal evaluation of the effect of the intervention and the manner in which it may be influenced by measured study characteristics and data quality. Data from the data-extraction spreadsheet is tabulated to form a summary of the number of data sets providing a yes, no, or neutral answer to each question (vote counting).

More formal quantitative analysis can be undertaken to generate overall point estimates of the effect size and to analyze reasons for heterogeneity in the effect of the intervention where appropriate data exist. Meta-analysis is now commonly used in ecology (e.g., Arnqvist & Wooster 1995; Osenberg et al. 1999; Gates 2002), so we have not treated it in detail here. Meta-analysis provides summary effect sizes with each data set weighted according to some measure of its importance, with more weight given to large studies with precise effect estimates and less to small studies with imprecise effect estimates. Generally each study is weighted in inverse proportion to the variance of its effect. Pooling of individual effects can be undertaken with fixed-effects or random-effects statistical models. Fixed-effects models estimate the average effect and assume there is a single, true underlying effect, whereas random-effects models assume there is a distribution of effects that depend on study characteristics. Random effects models include interstudy variability (assuming a normal distribution); thus, when there is heterogeneity, a random-effects model

has wider confidence intervals on its summary effect than a fixed-effect model. In medicine both statistical models are used to assess the robustness of statistical synthesis with an a priori decision about which is most germane (NHS CRD 2001; Khan et al. 2003). Results of our initial reviews suggest that random-effects models are most appropriate for the analysis of ecological data because the numerous complex interactions common in ecology are likely to result in heterogeneity between studies.

Relationships between differences in characteristics of individual studies and heterogeneity in results can be investigated as part of the meta-analysis, thus aiding the interpretation of ecological relevance of the findings. Exploration of these differences is facilitated by construction of tables that group studies with similar characteristics and outcomes together. Data sets can be stratified into subgroups based on populations, interventions, outcomes, and methodology. Important factors that could produce variation in effect size should be defined a priori (see stage 1 above) and their relative importance considered prior to data extraction to make the most efficient use of data. Differences in subgroups of studies can then be explored.

If sufficient data exist, meta-analysis can be undertaken on subgroups and the significance of differences assessed. Such analyses must be interpreted with caution because statistical power may be limited (Type I errors possible) and multiple analyses of numerous subgroups could result in spurious significance (Type II errors possible). Alternatively, a meta-regression approach can be adopted whereby linear regression models are fitted for each covariate, with studies weighted according to the precision of the estimate of treatment effect in a random-effects model (Sharp 1998).

Despite the attempt to achieve objectivity in reviewing scientific data, considerable subjective judgment is required when undertaking meta-analyses. These judgments include decisions about the choice of effect measure, how data are combined to form data sets, which data sets are relevant and which are methodologically sound enough to be included, methods of meta-analysis, and the issue of whether and how to investigate sources of heterogeneity (Thompson 1994). Reviewers should explicitly state and distinguish between the a priori and post hoc rationales behind these decisions to minimize bias and increase transparency.

A review of the impact of wind turbines on bird abundance utilized standardized mean difference meta-analysis with weighting by inverse variance to combine data from 19 globally distributed wind farms. Sensitivity analyses were used to explore the effect of including data from unreplicated studies and to assess bias arising from data extraction of pseudoreplicated or aggregated data. Pooled effect sizes remained negative and statistically significant regardless of how the effect sizes were generated, indicating that the patterns in the data were robust. A priori and post hoc reasons for heterogeneity were explored with meta-regression. Of the a priori variables only bird taxon appeared to modify the result, with relationships between turbine number and power being too weak to have biological significance. Post hoc analysis revealed that the impact of wind farms became more pronounced over time, a finding not reported by any of the original research or previously assessed in the literature. This has important implications because declines in local bird abundance are more likely to have deleterious population-level impacts if they worsen over time. It also suggests that current wind-farm monitoring programs are of inadequate duration to detect deleterious effects.

Stage 3—Reporting and Dissemination of Results

Before reports are disseminated they should be subjected to expert scrutiny or peer review, including assessment of scientific quality and completeness. This process requires the development of an editorial panel equivalent to that of a journal or grant board, but with a more supportive role in helping reviewers achieve the necessary quality rather than rejecting large numbers outright.

The recommended format for reporting is a short summary that highlights the main review outcomes. This should be written so as to enable effective communication with managers and policy formers. A full report, written for the commissioning body, and internal records will normally include too much detail for wider dissemination but should nevertheless be available, along with the summary, to all who want more information on the conduct of the review process. Commonly, the review will also be submitted, at the author's discretion, for publication in a peer-reviewed journal. We have developed separate guidelines and a format for presentation of reviews (www.cebc.bham.ac.uk/gettinginvolved.htm).

A full consideration of dissemination and implementation activities is beyond the scope of this paper, but a few general comments are pertinent. Wide dissemination and open access are key requirements of the evidence-based framework. However, standards of review have to be ensured; therefore, a central Web site administered by a collaboration of stakeholders is recommended, following the Cochrane Collaboration model with its emphasis on transparency of the review process and independence from bias (Fazey et al. 2004). On acceptance through peer review, summaries of reviews should be posted on the Web site with free access. Such a Web resource will be of limited use until many more systematic reviews have been undertaken.

Requirement for Further Work

To date, no systematic reviews have been published in ecology without involvement of the authors. There is

therefore potential for bias in development of appropriate methodology. For example, all reviews to date have incorporated comparators, although work in progress involves synthesizing experience and evidence with Bayesian methodologies (Morris & Normand 1992; Louis & Zelterman 1993). It could be argued that this is an excessively reductionist approach, applying a narrow definition of evidence (Fox 2005) and that further methodological development might be necessary to integrate different types of evidence (Dixon-Woods et al. 2004) or to assess ecological information of types beyond the experience of the authors.

Other issues require consideration to strengthen the ecological guidelines presented above. Medical systematic review methodology is developing rapidly, with new techniques being developed to handle the variable levels of data quality in fields such as diagnostic testing. The utility of these techniques for ecological purposes requires further investigation. Likewise, techniques for economic cost-benefit evaluation and disseminating evidence to different audiences (political, scientific, practitioner, and stakeholder groups) (NHS CRD 2001) warrant consideration. Addressing all these issues is beyond the scope of this paper, but they require further development if an ecological evidence base is to be fully established. The ecological guidelines presented evolved from the existing medical model. Table 3 highlights key differences between ecological and medical guidelines at present, but as experience with ecological systematic review grows, the guidelines should be revised and updated as is standard practice in medicine.

As was the experience in the medical field, it will take time for systematic reviews to be recognized and valued as equivalent to other scientific papers in conservation. Key steps forward in encouraging more systematic reviews will be for journals to encourage their submission and publication and for funders to see systematic reviews as a valid form of research. We call on the conservation and environmental management communities to engage with us to further develop the ecological systematic review and create the accessible evidence base that the subject urgently requires.

Table 3. Differences between the medical systematic review guidelines and the ecological review guidelines advocated by the authors.

Review stage	*Medical guidelines*	*Ecological guidelines*
Question formulation	question formulation generally not limited by complexity and study numbers	question formulation usually limited by information availability and complexity requiring a balance between holism (more realistic) and reductionism (more studies)
	stakeholder engagement useful but not generally critical	stakeholder engagement may be critical because conservation actions often result in conflicts in objectives
Developing review protocol: search strategy	complex searches balancing sensitivity and specificity are possible and recommended	high sensitivity, low specificity searches are recommended to reduce bias and increase repeatability because ecology lacks the sophisticated search infrastructure of medicine
Assessing quality of methodology	clear hierarchy of evidence generally applicable and often used to define a minimum quality threshold	pragmatic quality weightings and sensitivity analyses must augment data-quality hierarchies to avoid misinterpretation, particularly when combining data across the hierarchy to increase sample sizes
	performance bias and detection bias addressed by blinding; methodology is easy to assess with published quality weightings; and attrition bias is common	performance bias and detection bias addressed by experimental design but are hard to assess especially in a standardized manner, necessitating the use of review-specific quality weightings; attrition bias rare
	numerous off-the-shelf checklists available to assess the validity of medical research	no off-the-shelf checklists, hence the need for a priori review-specific criteria preferably validated by consensus with stakeholders
Data extraction	data extraction often relatively straightforward, except for missing data and data hygiene problems	data extraction complex especially with respect to variance measures for weighting; a priori rules must be developed in order to extract data in a repeatable, standardized manner; independence and (pseudo)replication are common problems
Data synthesis: meta-analysis	fixed and random effects models are applicable	random-effects models are generally more useful than fixed-effect models because the complex interactions in ecology generally result in ecologically important heterogeneity between studies

Acknowledgments

We thank the many conservation managers and scientists who have given us constructive feedback on the review process. Medical colleagues have contributed guidance particularly, T. Knight, R. Taylor, and K. Khan. We would also like to thank Ioan Fazey and two anonymous reviewers for their comments on an earlier draft of this work. This work was supported through grants from English Nature, the U.K. Natural Environment Research Council, and the Royal Society for the Protection of Birds.

Literature Cited

Arnqvist, G., and D. Wooster. 1995. Metaanalysis—synthesizing research findings in ecology and evolution. Trends in Ecology & Evolution **10:**236–240.

Bero, L., R. Grilli, J. Grimshaw, G. Mowatt, A. Oxman, and M. Zwarenstein, editors. 1999. Effective practice and organisation of care module of the Cochrane Database of Systematic Reviews. Issue 2. Update software. The Cochrane Library, Oxford, United Kingdom.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20:**37–46.

Cooper, H. M. 1984. Integrating research. A guide for literature reviews. Sage Publications, Newbury Park.

Dixon-Woods, M., S. Agarwal, B. Young. D. Jones, and A. Sutton. 2004. Integrative approaches to qualitative and quantitative evidence. National Health Services Health Development Agency, London.

Edwards, P., M. Clarke, C. DiGuiseppi, S. Pratap, I. Roberts, and R. Wentz. 2002. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Statistics in Medicine **21:**1635–1640.

Emerson, J. D., E. Burdick, D. C. Hoaglin, F. Mosteller, and T. C. Chalmers. 1990. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled Clinical Trials **11:**339–352.

Fazey, I., J. G. Salisbury, D. B. Lindenmayer, J. Maindonald, and R. Douglas. 2004. Can methods applied in medicine be used to summarize and disseminate conservation research? Environmental Conservation **31:**190–198.

Feinstein, A. R. 1985. Clinical epidemiology: the architecture of clinical research. Saunders, Philadelphia.

Feinstein, A. R., and R. I. Horwitz. 1982. Double standards, scientific methods, and epidemiological research. New England Journal of Medicine **307:**1611–1617.

Fox, D. M. 2005. Evidence of evidence-based health policy: the politics of systematic reviews in coverage decisions. Health Affairs **24:**114–122.

Gates, S. 2002. Review of methodology of quantitative reviews using meta-analysis in ecology. Journal of Animal Ecology **71:**547–557.

Gotzsche, P. C. 1987. Reference bias in reports of drug trials. British Medical Journal **295:**654–656.

Hedges, L. V. 1994. Statistical considerations. Pages 30–33 in H. Cooper and L.V. Hedges, editors. The handbook of research synthesis. Russell Sage Foundation, New York.

Higgins, J. P. T., and S. Green, editors. 2005. Cochrane handbook for systematic reviews of interventions 4.2.5. John Wiley & Sons, Chichester, United Kingdom.

Horwitz, R. I., and A. R. Feinstein. 1979. Methodological standards and contradictory results in case-control research. American Journal of Medicine **66:**556–564.

Jackson, G. B. 1980. Methods for integrative reviews. Review Education Research **50:**438–460.

Jüni, P., A. Witschi, R. Bloch, and M. Egger. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. Journal of American Medical Association **282:**1054–1060.

Khan, K. S., R. Kunz, J. Kleijnen, and G. Antes. 2003. Systematic reviews to support evidence-based medicine: how to apply findings of healthcare research. Royal Society of Medicine Press, London.

Kunz, R., and A. D. Oxman. 1998. The unpredictability paradox: review of empirical comparisons of randomised and nonrandomised trials. British Medical Journal **317:**1185–1190.

Leimu, R., and J. Koricheva. 2005. What determines the citation frequency of ecological papers? Trends in Ecology & Evolution **20:**28–32.

Levine, M., S. Walter, H. Lee, T. Haines, A. Holbrook, and V. Moyer. 1994. The Evidence-Based Medicine Working Group. Users' guides to the medical literature IV: how to use an article about harm. Journal of American Medical Association **271:**1615–1619.

Light, R. J., and D. B. Pillemer. 1984. Summing up: the science of reviewing research. Harvard University Press, Cambridge, Massachusetts.

Louis, T., and D. Zelterman. 1993. Bayesian approaches to research synthesis. Pages 411–422 in H. Cooper and L.V. Hedges, editors. The handbook of research synthesis. Russell Sage Foundation, New York.

Moher, D., A. R. Jadad, G. Nichol, M. Penman, P. Tugwell, and S. Walsh. 1995. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Controlled Clinical Trials **16:**62–73.

Moher, D., A. R. Jadad, and P. Tugwell. 1996. Assessing the quality of randomized controlled trials: current issues and future directions. International Journal of Technology Assessment in Health Care **12:**195–208.

Morris, C. N., and S. L. Normand. 1992. Hierarchical models for combining information and for meta-analyses. Pages 321–344 in J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Bayesian statistics. 4th edition. Oxford University Press, New York.

NHS (National Health Service) CRD (Centre for Reviews and Dissemination). 2001. Undertaking systematic review of research on effectiveness. NHS CRD, University of York, York, United Kingdom.

Osenberg, C. W., O. Sarnelle, S. D. Cooper, and R. D. Holt. 1999. Resolving ecological questions through meta-analysis: goals, metrics and models. Ecology **80:**1105–1117.

Pullin, A., and T. Knight. 2001. Effectiveness in conservation practice: pointers from medicine and public health. Conservation Biology **15:**50–54.

Pullin, A., and T. Knight. 2003. Support for decision-making in conservation practice: an evidence-based approach. Journal for Nature Conservation **11:**83–90.

Pullin, A., T. Knight, D. Stone, and K. Charman. 2004. Do conservation managers use scientific evidence to support their decision-making? Biological Conservation **119:**245–252.

Ravnskov, U. 1992. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. British Medical Journal **305:** 9–15.

Roberts, P.D., G. B. Stewart, and A.S. Pullin. 2006. Are review articles a reliable source of evidence to support conservation management? A comparison with medicine. Biological Conservation:in press.

Sharp, S. 1998. Meta-analysis regression: statistics, biostatistics, and epidemiology. Stata Technical Bulletin **42:**16–22.

Schulz, K. F., I. Chalmers, R. J. Hayes, and D. G. Altman. 1995. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the American Medical Association **273:**408–412.

Stevens, A., and R. Milne. 1997. The effectiveness revolution and public health. Pages 197–225 in G. Scally, editor. Progress in public health. Royal Society of Medicine Press, London.

Stewart, G. B., C. F. Coles, and A. S. Pullin. 2005*a*. Applying evidence-based practice in conservation management: lessons from the first systematic review and dissemination projects. Biological Conservation **126:**270–278.

Stewart, G. B., A. S. Pullin, and C. F. Coles. 2005*b*. Effects of wind turbines

on bird abundance. Systematic review 4. Centre for Evidence-Based Conservation, Birmingham, United Kingdom. (Also available from www.cebc.bham.ac.uk/systematicreviews.htm.)

Sutherland, W., A. Pullin, P. Dolman, and T. Knight. 2004. The need for evidence-based conservation. Trends in Ecology & Evolution **19:**305–308.

Thompson, S. 1994. Systematic review: why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal **309:**1351–1355.

Tyler, C., and A. S. Pullin. 2005. Do commonly used interventions effectively control *Rhododendron ponticum*? Systematic Review 6. Centre for Evidence-Based Conservation, Birmingham, United Kingdom. (Also available from www.cebc.bham.ac.uk/systematicreviews.htm.)

Tyler, C., E. Clark, and A. S. Pullin. 2005. Do management interventions effectively reduce or eradicate populations of the American Mink, *Mustela vison*? Systematic Review 7. Centre for Evidence-Based Conservation, Birmingham, United Kingdom. (Also available from www.cebc.bham.ac.uk/systematicreviews.htm.)

Tyler, C., A. S. Pullin, and G. B. Stewart. 2006. Effectiveness of management interventions to control invasion by *Rhododendron ponticum*. Environmental Management **37:**513–522.