Methods Research Report

Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract Nos.: 290-2007-10055-I (Tufts EPC), 290-2007-10059-I (University of Ottawa EPC), 290-2007-10062-I (Southern California EPC), 290-2007-10063-I (ECRI Institute EPC)

Prepared by:
Tufts Evidence-based Practice Center, Tufts Medical Center
Boston, MA
with input and assistance from
University of Ottawa Evidence-based Practice Center, University of Ottawa
Ottawa, ON
Southern California Evidence-based Practice Center
Santa Monica, CA
ECRI Institute Evidence-based Practice Center
Plymouth Meeting, PA

Investigators
Ethan M. Balk, M.D., M.P.H.
Mei Chung, Ph.D.
Minghua L. Chen, M.D., M.P.H.
Thomas A. Trikalinos, M.D., Ph.D.
Lina Kong Win Chang, B.S.

AHRQ Publication No. 12(13)-EHC145-EF
January 2013

This report is based on research conducted by the Tufts Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10055-I). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials that are clearly noted in the document. Further reproduction of those copyrighted materials is prohibited without the specific permission of copyright holders.

Persons using assistive technology may not be able to fully access information in this report. For assistance, contact EffectiveHealthCare@ahrq.hhs.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Balk EM, Chung M, Chen ML, Trikalinos TA, Kong Win Chang L. Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages. Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-1.) Rockville, MD: Agency for Healthcare Research and Quality. January 2013. AHRQ Publication No. 12(13)-EHC145-EF. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by email to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang M.D., M.P.H.
Director, EPC Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Kim Marie Wittenberg, M.A.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

We would like to thank and acknowledge the following investigators for their insights into the study design and methods and/or for their data extractions: Sophia Tsouros, B.H.Kin, University of Ottawa Evidence-based Practice Center; Mohammed T. Ansari, M.B.B.S., M.Med.Sc., M.Phil., University of Ottawa Evidence-based Practice Center; Jordi Pardo Pardo, University of Ottawa, Centre for Global Health; Sydne J. Newberry, Ph.D., Southern California Evidence-based Practice Center; Susanne Hempel, Ph.D., Southern California Evidence-based Practice Center; Melisa Chang, M.D., Southern California Evidence-based Practice Center; Paul G. Shekelle, M.D., Ph.D., Southern California Evidence-based Practice Center; Joann Fontanarosa, Ph.D., ECRI Institute Evidence-based Practice Center; Fang Sun, M.D., Ph.D., ECRI Institute Evidence-based Practice Center; Katrin Uhlig, M.D., M.S., Tufts Evidence-based Practice Center; Teruhiko Terasawa, M.D., Ph.D., Fujita Health University, School of Medicine, Tsu, Japan; Nikolaos A. Trikalinos, M.D., University of Maryland, Marlene and Stewart Greenebaum Cancer Center, Baltimore, MD; Fumiaki Imamura, Ph.D., Harvard School of Public Health, Boston, MA; and Issa J. Dahabreh, M.D. M.S., Tufts Evidence-based Practice Center. Boston, MA.

Assessing the Accuracy of Google Translate To Allow Data Extraction From Trials Published in Non-English Languages

Structured Abstract

Background: One of the strengths of systematic reviews is that they aim to include all relevant evidence. However, study eligibility is often restricted to the English language for practical reasons. Google Translate, a free Web-based resource for translation, has recently become available. However, it is unclear whether its translation accuracy is sufficient for systematic reviews. An earlier pilot study provided some evidence that data extraction from translated articles may be adequate but varies by language. To address several limitations of the pilot study, four collaborating Evidence-based Practice Centers conducted a more rigorous analysis of translations of articles from five languages.

Methods: We included 10 randomized controlled trials in 5 languages (Chinese, French, German, Japanese, and Spanish). Eligible studies were trials that reported per-treatment group results data. Each article was translated into English using Google Translate. The time required to translate each study was tracked. The original language versions of the articles were double data extracted by fluent speakers and reconciled. Each English-translated article was extracted by two of eight researchers who did not speak the given language. These 8 researchers also each extracted 10 English-language trials to serve as a control. Data extracted included: eligibility criteria, study design features, outcomes reported, intervention and outcome descriptions, and results data for one continuous and/or one categorical outcome. We used a generalized linear mixed model to examine whether the probability of correctly extracting an item from a translated article is related to the language of original publication. The model used each extractor's accuracy in extracting the English language trials to control for reviewer effects.

Results: The length of time required to translate articles ranged from 5 minutes to about 1 hour for almost all articles, with an average of about 30 minutes. Extractors estimated that most Spanish articles required less than 5 additional minutes to extract because of translation issues, but about two-thirds of other language articles required between 6 and 30 additional minutes for extraction. Analyses of the adjusted percentage of correct extractions across items and languages and of the adjusted odds ratio of correct extractions compared with English revealed that in general, across languages the likelihood of correct extractions was greater for study design and intervention domain items than for outcome descriptions and, particularly, study results. Translated Chinese articles yielded the highest percentage of items (22 percent) that were incorrectly extracted more than half the time (but also the largest percentage of items, 41 percent, that were extracted correctly more than 98 percent of the time. Relative to English, extractions of translated Spanish articles were most accurate compared with other translated languages.

Conclusion: Translation generally required few resources. Across all languages, data extraction from translated articles was less accurate than from English language articles, particularly and importantly for results data. Extraction was most accurate from translated Spanish articles and least accurate from translated Chinese articles. Use of Google Translate has the potential of being an approach to reduce language bias; however, reviewers may need to be more cautious

about using data from these translated articles. There remains a tradeoff between completeness of systematic reviews (including all available studies) and risk of error (due to poor translation).

Contents

Introduction

Systematic reviews conducted by the Agency for Healthcare Research and Quality (AHRQ) Evidence-based Practice Centers (EPCs) most commonly restrict literature searches to English language publications. In a sample of 10 recent Evidence Reports (numbers 189-198), 8 were restricted to English-language publications. One report included studies in languages for which the EPC had "available fluency" and only one reported not restricting by language. Among 28 other recent Comparative Effectiveness Reviews (CERs) with final or draft documents downloadable from the AHRQ Web site, 20 were restricted to English-language publications. Four explicitly did not impose any language restriction. Two did not report language restriction in their methods chapter and included one study each in Dutch and German. One placed no language restriction on comparative studies but included only English-language cohort studies. One included German- and French-language studies for nonoperative interventions (which were sparse), but only English-language publications for operative treatments "due to lack of translation resources." Three of the CERs wrote that the language restriction was due to lack of resources or prohibitive translation costs, despite the recognition in one CER "that requiring studies to be published in English could lead to bias."

Thus, in most instances, EPC reports may be at risk of selection bias based on language (if there is reason to suspect differential publication of studies in English language and non-English journals)1 and may not be following Standard 3.2.6 from the recent Institute of Medicine's (IOM) "Finding What Works in Health Care: Standards for Systematic Reviews,"[2] "Search for studies reported in languages other than English if appropriate." The IOM report notes that there is some known evidence of language bias (e.g., investigators in Germany may be more likely to publish their negative results in German language publications and their positive results in English language publications).[1,3] However, numerous other studies have found that excluding non-English publications may not result in substantial bias (changes in estimates of treatment effects).[4-10] Nevertheless, excluding studies solely based on language runs counter to the concept of systematic review, of including all known evidence, particularly as investigators are being encouraged to include non-peer-reviewed and other studies in the grey literature.

Using a literature search module for randomized controlled trials,[11] a search in Medline from 1996 to May 25, 2012, found that of 2,982,047 citations, 92 percent were published in English. Table 1 shows the number and frequency of publications in other languages with more than 0.5 percent penetration.

Table 1. Percentage of studies from Medline search for randomized controlled trials in various languages

Language	N	Percent
Total (1996 – May 25, 2012)	2,982,047	100.00%
English	2,739,141	91.85%
Chinese	50,849	1.71%
German	39,170	1.31%
Russian	34,258	1.15%
French	29,287	0.98%
Spanish	27,049	0.91%
Japanese	16,915	0.57%

EPCs have varying capacities to extract non–English-language articles, based on the language knowledge of their staff. Formally translating all non–English-language articles is costly and resource-intensive, particularly if performed at the stage of full-text article screening. Therefore, a reliable, free, easily available service to translate articles may allow EPCs to easily broaden the scope of their systematic reviews, without introducing possible language bias by restrictions based on language. Google Translate® is a free, Web-based program with an excellent reputation for accurate, natural translation (http://translate.google.com). It is one of several such tools, including Yahoo!® Babel Fish (www.babelfish.com/), SDL FreeTranslation® (www.freetranslation.com), and Bing® Translator (www.bing.com/translator). In an analysis of four translation tools for a limited set of language pairs, Google Translate was found to perform best based on human judgment of translation accuracy.[12] A subsequent study comparing 2,550 language pairs (51 languages) in Google Translate using an automated technique to compare translations found a range of translation accuracy and that "translations between European languages are usually good, while those involving Asian languages are often relatively poor. Further, the vast majority of language combinations probably provide sufficient accuracy for reading comprehension in college."[13] Also of note, a pilot study presented as a poster at the 2009 Singapore Cochrane Collaboration meeting used Google Translate on 11 German articles from one Cochrane review and found that interrater agreement was 73 percent (κ=0.38) for whether the article should be included in the review.[14]

Tufts EPC recently conducted a pilot study evaluating Google Translate for data extraction from 88 articles published in 9 languages (Chinese, French, German, Hebrew, Italian, Japanese, Korean, Portuguese, and Spanish).[15] Briefly, the results of the study concluded that the length of time required to translate articles ranged from seconds (51 articles, 58 percent) to about 1 hour. Assessment by those who extracted the 88 translated articles indicated that "a little" extra time was required for 40 articles (45 percent) and "a lot" for 42 (48 percent). When evaluating all extraction items together, Portuguese and German articles had the best agreement between original and translated extractions, with high agreement between extractors among about 60 percent of the items, compared with 80 percent in English articles. Spanish, Hebrew, and Chinese had the lowest agreement (30, 24, and 8 percent, respectively). The absolute agreement and the proportion of items with high agreement were statistically significantly worse for all languages, compared with English. Eight of 10 English-language articles had high agreement for all items; compared with 7 of 10 Portuguese articles; 6 of 10 German articles; 4 of 10 French, Italian, and Korean; 3 of 8 Hebrew articles; 3 of 10 Japanese and Spanish articles; but no Chinese articles. However, the pilot study had several important limitations, including that only single extractions were performed of the native language articles and confirmation could not be conducted; the analyses did not allow for full differentiation between disagreements in extractions due to poor translation or due to different extractors interpreting articles in different ways or errors in extraction.

Aims

The current study was designed to form a collaboration of EPCs to better analyze the accuracy of the freely available, online, translation tool—Google Translate—for the purposes of data extraction of articles in selected non-English languages. The collaboration allowed for double data extraction and a better consensus determination of the important extraction items to assess; we also implemented an improved analytic technique.

The research had the following aims:

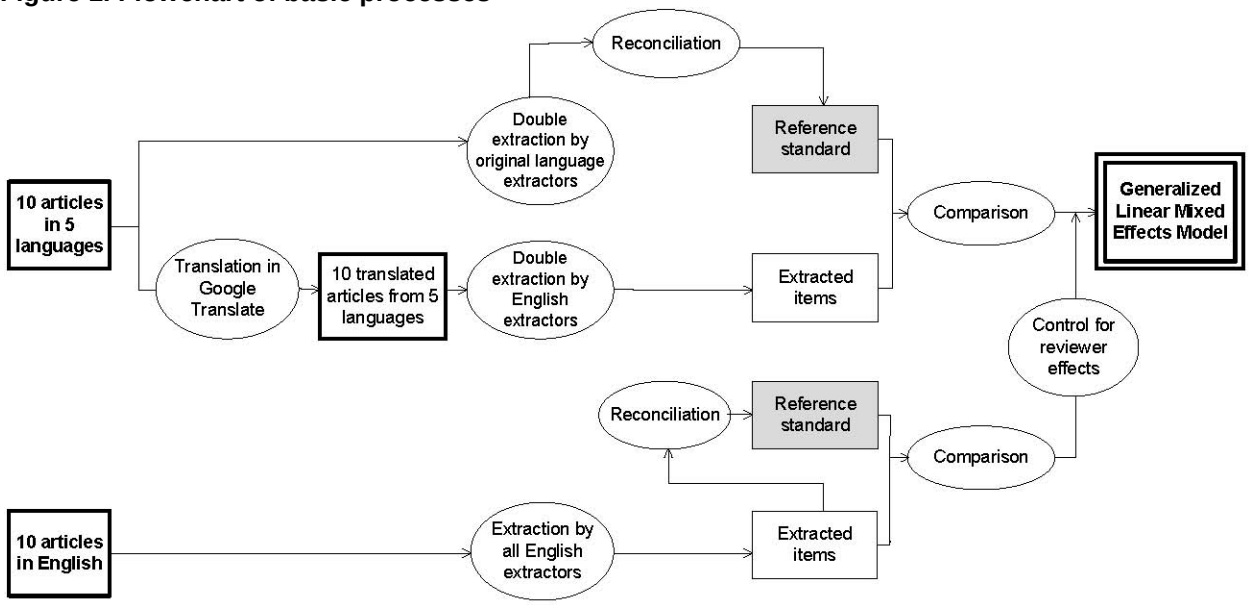
1. Compare data extraction of trials done on original-language articles by native speakers with data extraction done on articles translated to English by Google Translate.
2. Track and enumerate the time and resources used for article translation and the extra time and resources required for data extraction related to use of translated articles.

Methods

An invitation was sent to all EPCs to join in the current methods project. The invitation requested that participating EPCs nominate investigators to extract original language articles, translated articles, or both. For extraction from original language articles, EPCs were asked to offer fluent speaking investigators from within their EPCs, from their regular cohort of collaborating investigators, or from other investigators who could subcontract with them.

Figure 1 displays a flowchart of the basic processes of the data extraction, reconciliation, and analysis.

Figure 1. Flowchart of basic processes



Study Selection

Based on the frequency of non–English-language publications, a consensus on the most important languages to test, and the languages spoken by native speakers available to the collaborating EPCs, we included articles in the following five languages: Chinese, French, German, Japanese, and Spanish.

Using QUOSA Information Manager™ (v 8.07.265, QUOSA, Inc.) software, which allowed us to search in PubMed and automatically retrieve available PDF files, we searched with the term "randomized controlled trial," restricted separately to each of the five languages. This tool can retrieve PDF files from all journals for which the Tufts University Health Services Library has a subscription or that are publicly available. We accepted the first 10 publications of trials in each language, regardless of topic, for which either a machine readable PDF or HTML file was available for the full text of the article. We did not use a publication date limitation, but in practice we included more recent publications (since we accepted the first 10 eligible articles and articles had to be available in PDF or HTML file formats). We accepted only studies with these file types since otherwise they could not be translated with Google Translate.

We estimated that we had resources for approximately 10 articles per language. We explored whether this would be sufficient to show statistical significance for differences between languages as large as those observed in the pilot study between the language with the worst

agreement (50 percent among all items and extractors) and that with the best agreement (82 percent among all items and extractors).[15] We simulated 500 datasets for 10 reviewers extracting 10 articles each, for 5 (non-English) languages, assuming that reviewers were completely exchangeable and allocated in a balanced way; that papers were indistinguishable given language; and that the true probability for correct extraction in the five languages was regularly spaced between 50 and 82 percent. We ran logistic regression models similar to the ones used in the final analysis, and focused on the omnibus P-value for language factors (which expresses the language effect). The observed power to detect differences was above 80 percent.

Full-text articles were screened by a researcher who was native in that language to determine eligibility. Eligible studies were randomized controlled trials that reported per-treatment group results data. We excluded publications that had a simultaneous English translation in the PDF or HTML file. We also excluded publications that were not primary reports of trials (but were summaries of English-language trials).

In addition, we chose 10 English-language randomized controlled trials to use as a reference standard. Also using the same search technique in QUOSA, we arbitrarily chose English-language articles that met criteria and were published in a distribution of years roughly corresponding to the distribution of the non-English articles.

Translation

Each article was translated into English using Google Translate. This was done with the simplest method possible for each PDF (or HTML) file. Depending on the format of the articles, the English translations included the original tables and figures, translated the best they could be. We also copied over any English language abstracts that were published with the original articles, together with English language tables and figures. Each article was translated into a separate Word file that could be accessed without seeing the original article. Images of figures and tables that could not be translated due to formatting issues were included in the translated article files. The detailed methods for article translation are presented in the report of the pilot study.[15] Translations were performed primarily by a research assistant. A rough estimate of the time required to extract each study was tracked.

Data Extraction

All data extraction was performed in the data extraction tool, Systematic Review Data Repository (SRDR, Tufts Medical Center, Boston, Massachusetts), being developed at Tufts EPC.[16] The tool allows creation of flexible data extraction forms, direct entry of extracted data, reconciliation, and export of extracted data in an online format. The tool maintained independence of extractions by restricting extractors' access to their own extractions (except for administrators who had access to all extraction forms). The extraction items are listed in Appendix A.

Each original language version of the articles was double extracted by two fluent readers of medical text in those languages. The double extractions were checked for agreement at Tufts EPC, who informed the pairs of extractors of items that required reconciliation. The pairs of extractors (who were generally in different time zones) communicated by email and/or telephone to come to agreement. The reconciled version of the extractions from the original language articles served as the reference standard for those articles.

The translated versions of the articles, which included the English language abstract when available, were extracted by eight researchers who did not speak the original language of the

articles. Each translation was extracted by two researchers. Assignments were made arbitrarily (though not strictly randomly) with the goals of distributing extractors across languages, avoiding frequent pairing of researchers, and avoiding assigning translated articles to researchers who could read the original language of the articles. These researchers also extracted the 10 English-language articles (to serve as a control for "reviewer effects"; see Analysis below). Reconciliation of the extractions of English language articles was conducted by consensus either of five of eight extractors or, failing that, agreement between the two senior researchers at Tufts EPC.

Data Extraction Form

Since we were primarily interested in the accuracy of the data extraction, as opposed to the accuracy of all the text, we performed limited data extraction focusing on those study features that are most important for assessing the study characteristics, methods, and results (see Appendix A for the data extraction items).

We extracted the following information: inclusion and exclusion criteria; funding source; number of study centers; followup duration; whether the article reported randomization technique, allocation concealment method, intention-to-treat analyses, a power calculation; blinding (subjects, caregivers, outcome assessors, double blinding, single blinding); outcomes reported (see last paragraph in this section); interventions and controls, and their dose, frequency, route, and duration; numbers randomized to each intervention; and outcome descriptions. For a single (preselected) continuous outcome, we extracted the number of subjects analyzed per intervention, whether mean or median data were reported, the net difference (or difference between final values, depending on reported data), its standard error, the reported P value for the difference between groups, and what, if any, factors were adjusted for. For a single (preselected) categorical (dichotomous) outcome, we extracted the number of events (counts) and number analyzed per intervention, the odds ratio with its 95 percent confidence interval, the reported P value for the difference between groups, and what, if any, factors were adjusted for. We provided extractors with a standardized calculator in Excel for use when between-group comparison data were not reported but needed to be calculated (e.g., for odds ratio). In addition, we also asked extractors of translated articles two subjective, best guess questions: to provide a rough estimation of how much extra time they believe they spent with the article compared to an extraction of a similar English article (<5 minutes extra time, 6 to 30 minutes extra time, or >30 minutes extra time), their level of confidence in the accuracy and translation of the article (little confidence, moderate confidence, or strong confidence). Extractors were also asked for examples of poor translations that made extraction difficult.

Upon analyzing the extracted data, the following items were removed from analysis: whether analyses were adjusted and what the analyses were adjusted for (extremely few articles had adjusted analyses extracted); and calculated odds ratio (which was not consistently calculated correctly and was redundant with the counts and numbers analyzed).

Whenever possible, we selected one categorical outcome and one continuous outcome from each trial. Ideally, one of the two outcomes was presented in the abstract (and the full text) and one was presented in full text only. When necessary, we limited the extraction of results to two of multiple interventions, with a preference for the intervention most similar to a "control" and another arbitrary intervention. Also when necessary, a single timepoint was arbitrarily chosen, with a preference for longer followup periods.

Prior to data extraction, for each language we compiled a list of about a dozen outcomes that were reported in at least one article in that language. This was done with the assistance of native speakers. We aimed for a mix of primary and secondary outcomes, and clinical and intermediate (or surrogate) outcomes. During data extraction, researchers were asked to review the list of outcomes and check off all outcomes that were reported in the article. This item was added with the goal of determining the accuracy of finding potential outcomes of interest in translated articles (which would determine whether a study is included for a particular outcome).

Data Extraction Comparison

Reconciliation of the original language extractions (including English) and then comparisons of translations with the reconciled original language extractions were conducted either within SRDR or with data exported into Excel. For the comparisons of translations (and of English extractions) with their reference standards, each data item was coded as agree or disagree. "Disagree" included erroneous data, incomplete data (e.g., descriptions of eligibility criteria, lists of outcomes reported), and data items incorrectly extracted as not reported (no data).

For numerical data and for questions that allowed checkboxes or pulldown menus, comparisons were simple and objective. However, several questions involved text responses that required subjective assessment to determine agreement, including inclusion criteria, exclusion criteria, and descriptions of many interventions that were not drugs and required more qualitative descriptions and did not have a "dose," "frequency," or "route." For calculated numerical data, we were liberal in determining agreement, accepting mean differences, standard errors, odds ratios, and confidence intervals that were similar as being in agreement. We did this to minimize counting calculation differences as disagreements due to translation. During reconciliation of the English article extractions we found that for numerous articles there were reasonable differences in interpretation that resulted in different data being extracted. For example, vague wording led to different interpretations as to how many participants were analyzed or whether the reported P value was the P-value of the difference between interventions that we were interested in. Another common example was how to handle descriptions of interventions (e.g., how to handle the dose of a topical application where as much is used as is needed [no data vs. the concentration of the lidocaine in the application], the dose of therapy sessions [30-minute sessions vs. no data]). Since we could not adequately evaluate for all extractions of all articles whether differences were due to different interpretations or errors, we changed our approach to reconciliation. Where there were discrepancies between the double data extractors (from the original language articles) we asked the extractors to recheck and confirm their answers. If discrepancies remained after data checking, then both answers were accepted as correct answers. If the translated article extractor extracted either of the "correct" answers, this was treated as an agreement.

Analysis

All 8 reviewers each extracted all 10 English-language papers, which served as a "common reference" in the analysis. Each translated non–English-language paper was extracted by two of the eight reviewers. It was not possible to allocate reviewers to non–English-language papers in a balanced way, because some reviewers were fluent in some non-English languages and for logistic reasons.

For each item, we used a generalized linear mixed effects model to examine whether the probability of correctly extracting the item is related to the language of the original publication and to the reviewer (extractor), accounting for the fact that reviewer extractions are grouped by

paper. The model used the pattern of allocation of reviewers to languages to control for reviewer effects, and is a logistic regression with fixed slope and random intercept terms. Specifically, the logit-transformed probability of correctly extracting an item from paper *i* written in original language *l* is modeled as a linear function:

$$\text{logit}(P_{correct}) = \alpha_0 + \sum_{r=2\ldots8} \beta_r I(reviewer=r) + \sum_{r=2\ldots6} \gamma_r I(language=l) + \zeta_{li},$$

where α_0 is an intercept, $I(\ )$ is the indicator function, $r=1,\ldots,8$ indexes reviewers, $l=1,\ldots,6$ indexes original languages, and ζ_{li} are paper-specific random intercepts normally distributed around 0. Reviewer 1 and language 1 (English) are the reference indicators. In the model, the βs are log odds ratios corresponding to reviewer effects. Similarly, the γs are the log odds ratios corresponding to language effects. The αs are random intercepts.

For each item, we report odds ratios for correct extraction for papers originally published in languages other than English compared with English-language papers. We use the fitted model to derive the predicted probability of correct classification for a hypothetical reviewer who is in some sense an average of the eight reviewers who extracted translated data. For some items virtually all extractions were correct, and the above model (and its fixed effects version) did not converge (perfect prediction). In these cases we ignored reviewer effects and calculated "crude" odds ratios for languages, and crude predicted probabilities of correct classification per language.

Separate from the main analyses, we used kappa statistics to quantify between-rater agreement in classifying five items on study risk of bias (methodological quality): randomization method, allocation concealment method, patient and caregiver blinding, outcome assessor blinding, and attrition rate (see Appendix A). We compared the kappa for agreement among the reviewers who extracted information from the original papers, with the kappa for agreement among reviewers who used the translated papers. We performed sensitivity analyses by merging "unclear" with "low" and "unclear" with "high" risk of bias. We used the Landis and Koch interpretation of values of kappa to determine the level of agreement.[17]

κ	Interpretation
< 0	Poor agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Results

The articles were chosen by language only. We did not consider geographic distribution when selecting articles. All Chinese, German, and Japanese articles were from China, Germany, and Japan, respectively. The French articles were from France (5), Canada (3), Tunisia (1), and Turkey (1); and the Spanish articles were from Spain (5), Argentina (3), Colombia (1), and Mexico (1). The Chinese articles were all published in simplified Chinese. Other characteristics of the included studies are presented in Table 2.

Table 2. Characteristics of included trials

Language	Publication Dates (N)	Clinical Domains / Populations	Intervention Types (N)	Outcome Types (N)
Chinese	2004 (1) 2008 (3) 2009 (1) 2010 (4) 2011 (1)	Colonoscopy Dementia Gynecologic surgery Hepatocellular carcinoma Lung cancer Myopia Nocturia Parkinson's Disease Premature LBW infants Pulmonary surgery	Behavior (1) CAM (4) Dxic test agent (1) Drug (4)	Dichotomous (8) Continuous (9)
French	1993 (1) 1994 (1) 1996 (1) 1997 (1) 2000 (1) 2006 (2) 2008 (2) 2009 (1)	Acute sinusitis Allergic conjunctivitis Helicobacter pylori infection Hepatitis C Homeopathy adverse effect Hypercholesterolemia Hypertension Obstetrics Pterygium (ophthalmology) Refraction (ophthalmology)	Counseling (1) CAM (1) Diet (1) Drug (7)	Dichotomous (7) Continuous (8)
German	2002 (1) 2005 (1) 2007 (2) 2008 (2) 2009 (1) 2010 (2) 2012 (1)	Anesthesia Cataract surgery Chronic prostatitis Hernia surgery Keratoplasty (ophthalmology) Knee arthroscopic surgery Scar formation Schizophrenia Total hip replacement Whiplash	CAM (1) Counseling (1) Device (3) Drug (3) Exercise (1) Surgery (1)	Dichotomous (7) Continuous (10)

Table 2. Characteristics of included trials (continued)

Language	Publication Dates (N)	Clinical Domains / Populations	Intervention Types (N)	Outcome Types (N)
Japanese	2001 (2) 2002 (1) 2004 (2) 2008 (1) 2009 (2) 2010 (2)	Cardiac function Colorectal cancer Diabetes mellitus education Diabetes mellitus prevention Fungal prophylaxis (oncology), 2 studies Gynecologic laparoscopic surgery Hypercholesterolemia Smoking cessation Tinea pedis	CAM (1) Counseling (1) Drug (4) Education (3) Formulation (1)	Dichotomous (5) Continuous (7)
Spanish	2002 (1) 2003 (1) 2005 (1) 2006 (1) 2008 (1) 2009 (2) 2010 (3)	Gerontology, enteral feedings Hypertension Intracranial hypertension Malaria Molar extraction (dental) Neonatology, 2 studies Obesity, 2 studies Oral candidiasis	Anesthesia (1) Drug (5) Education (1) Nutrition (2) Procedure (1)	Dichotomous (5) Continuous (7)
English	1997 (1) 2002 (1) 2003 (1) 2005 (1) 2007 (1) 2008 (1) 2009 (1) 2010 (2) 2011 (1)	Anesthesia, bowel surgery Cardiovascular risk factors Cerebral ischemia Cleft lip and palate Diabetes mellitus, type 2 Macular degeneration Menopause Nutrition, micronutrients Parkinson's disease Sickle cell anemia	Blood products (1) Drugs (4) Exercise (1) Nutrition (3) Procedure (1)	Dichotomous (6) Continuous (9)

CAM = complementary and alternative medicine; Dxic = diagnostic; LBW = low birth weight; N = number of articles.

Among the 15 investigators who extracted data from original language articles (10 investigators) and translated and English articles (9 investigators), 11 are M.D. or Ph.D. (or both) researchers, 2 are research associates (or equivalent), and 2 are medical residents with research experience. The median duration of experience with data extraction was 5 years, with 4 extractors having 10 to 14 years of experience, and 4 having less than 1 year of experience. Six of the investigators have participated in more than 20 systematic reviews, 1 has participated in 11 to 20 reviews, 5 in 6 to 10 reviews, and 3 in 5 or fewer reviews. Eight investigators have extracted more than 100 studies, 4 have extracted 51 to 100 studies, and 3 have extracted 50 or fewer articles. Nine investigators judged that they have a lot of comfort with the Cochrane risk of bias questions, 4 had moderate comfort, 1 had little comfort, and 1 had no experience with assessing risk of bias. The medical resident with no prior systematic review experience extracted 10 original language articles. She had oversight and assistance from the director of the EPC she was affiliated with.

Article Translation

The length of time required to translate articles ranged from 5 minutes (2 of 50 articles) to about 1 hour (11 articles) for most articles; 2 articles took more than 1 hour. Excluding the time taken for the latter two articles, the average time to translate was about 30 minutes. The time-for-translation distributions varied by language (Table 3), with Spanish articles being the quickest to translate and Chinese articles taking longest.

Table 3. Translation time (minutes), by language

Articles*:	1	2	3	4	5	6	7	8	9	10	Median
European											
French	20	30	30	30	30	30	30	30	45	60	30
German	15	20	20	25	30	30	30	30	40	240	30
Spanish	10	10	10	10	15	15	15	15	20	20	15
Asian											
Chinese	60	60	60	60	60	60	60	60	60	120	60
Japanese	5	5	20	20	20	20	30	30	30	60	20

*For each language, the approximate duration of time, in minutes, for translation of each article is listed, sorted from shortest to longest time.

In general, the European- and Japanese-language articles could be translated automatically from their PDF or HTML files. These texts were then copied to Word documents after translation. However, the ease of translation was largely related to the file and text types used by the journals and whether Google Translate could read these directly or not.

The extra time required to translate the other articles consisted mainly of iteratively copying blocks of text (paragraphs or columns) from the article into the Google Translate Web site and then copying the translated text into Word documents. This often involved using the appropriate alphabet from the original language, and removing false line breaks, hyphens, and unnecessary spaces. We discovered (and were informed by the Chinese speakers among us) that we needed to remove false line breaks (artifactual breaks not at the end of sentences) in the Asian language articles to allow meaningful translation. Translation of tables was frequently very time consuming as it required a large number of translations of individual row and column headers and formatting in the translated Word document.

For numerous articles, particularly those in the Chinese language, Google Translate could directly translate the PDF or HTML file, but the resulting file was unreadable because of heavily overlapping text across columns; therefore, manual copying and pasting of these articles had to be done. Since Google Translate attempted to maintain the original formatting and because some written languages are much more compact than English, the English text ran from one column to the next, overlapping the text in the second column.

Other issues we encountered included that one Spanish PDF could not be read originally but could after it was saved as a TIFF file from which another PDF was created; one German article required removing multiple instances of "¬" (an optional hyphen) before translation could succeed. One German article was clearly an outlier in that it took almost 4 hours to translate because of the poor quality of the original file. When text from this particular article was copied and pasted into either Google Translate or Microsoft Word, the copied text included spaces randomly placed within most of the words. Because the quality of the translated text was greatly improved after removing these superfluous spaces, this extra step was undertaken. One Chinese article took almost 2 hours because the non-Chinese characters (such as words and numbers) within the file were copied to gibberish and had to be manually retyped for proper translation.

Data Extraction From Translated Articles

The assessment by the English language data extractors was that extraction from translated articles generally took more time than extraction from an equivalent English-language article would have taken. Extractors were asked to estimate how much additional time they spent on each translated article compared with the time they likely would have spent with a comparable English article (Table 4). For Spanish articles, extractors estimated that 56 percent of articles

took less than 5 additional minutes to extract, and all but one article took up to 30 additional minutes to extract. Extraction of other translated articles generally took longer. Between 60 and 75 percent of other language articles were estimated to take between 6 and 30 additional minutes to extract, with generally most of the remaining articles requiring more than 30 minutes extra. Anecdotally, for some Chinese articles the translation was so poor that little could be extracted, which resulted in little time being required to extract the article.

Table 4. Estimated additional time required compared to extraction of a similar English article

Extra Time	Chinese Percent (n)	French Percent (n)	German Percent (n)	Japanese Percent (n)	Spanish Percent (n)	Overall Percent (n)
<5 min	20% (4)	5% (1)	20% (4)	5% (1)	56% (10)	21% (20)
6-30 min	70% (14)	68% (13)	60% (12)	75% (15)	39% (7)	63% (61)
>30 min	10% (2)	26% (5)	20% (4)	20% (4)	6% (1)	16% (16)

Extractors were also asked to assess their level of confidence in the translation of the articles (Table 5). Extractors had strong confidence for the majority (60 percent) of Spanish articles. Confidence in the translation of other language articles was generally moderate with 65 to about 70 percent of articles across languages.

Table 5. Confidence in accuracy and completeness of the translation

Confidence	Chinese Percent (n)	French* Percent (n)	German Percent (n)	Japanese Percent (n)	Spanish Percent (n)	Overall* Percent (n)
Strong	10% (2)	5% (1)	15% (3)	5% (1)	60% (12)	26% (26)
Moderate	65% (13)	65% (13)	65% (13)	60% (12)	25% (5)	55% (55)
Little	25% (5)	25% (5)	20% (4)	35% (7)	15% (3)	18% (18)

* 1 extractor did not rate confidence level for 1 article.

Table 6 provides some examples of unintelligible translations collected by the extractors. Additional issues included lack of translation of figures and some tables, blocks of gibberish, and completely untranslated text.

Table 6. Examples of poor translation, by language

Language	Translated Text
Chinese	R group of 33 cases, due to the three cases of epidural catheter prolapse were excluded, the remaining 18 cases analgesia is effective, 12 cases of analgesia is not valid Of CO2 group in check after 1, 3 and 6h abdominal pain mean VAS scores were significantly lower than the air group (all P <0.01, Figure 2), check the inter-24h two groups remain statistically different (P <0.05, Figure 2). stabbed five tomato feed rate of 2 per week <8mL/kg the number of night enuresis ≥ 1 or ≥ 7 times / week the main points to take Baihui, the Court of God, Ojo. Acupoints with the disease: liver and kidney deficiency with liver Yu, Shenshu; phlegm blocking orifices with Zhong Wan, Hong Leong; deficiency with the sea air. PD duration â‰¤ 15 years. 2009 in my inpatient children of low birth weight (less than 37 weeks gestational age at birth and neonatal birth weight less than 2500g) as the object of study, except for the digestive tract abnormalities, in line with the conditions60 cases were randomly divided into [...] Probiotic treatment groups at the beginning of the gastrointestinal tract in children, while giving the Golden Bifid feeding treatment, each 0.25g, 2 times / day, oral or nasal administration, to the increase of milk when total parenteral nutrition is discontinued , premature children total parenteral nutrition and fluid requirements with reference to calorie "practical neonatology" (3rd edition) in the recommended amount
French	I1 important to note that this assessment does not address the practice of individual physicians each of the media but rather on an overall profile of theirapproach and more specifically 1'HTA on the proportion of hypertensive patients who have reached a satisfactory control their hypertension. a single pregnancy by presenting the evolutionary eutrophic tion of the top and run over 36 weeks amenorrhea (SA) plus zero days treatment may interfere in the convenience store (anti-depressant) , rather than .
German	A patient with intrastromal corneal ring was excluded from the study because he suffered by a missile blast with Fadenruptur graft and bending of the corneal ring. auch here should be 3 – diclofenac 50 It was only to patients, each of two same age, in terms of localization, incision depth and the used suture scars showed equivalent or at least had a 20 cm long scar and Pa ¬, who stood so close to release that no meeting date has been more identified. standard therapy using HWSKrawatte The randomization of patients was based on randomisation. "instead gehabter fracture" Intervention" and was "Shamgruppe
Japanese	Terubinafinkurimu hydrochloride (A): 1g of terbinafine hydrochloride cream containing 10mg. Cream base (B): cream base only brewed Results Fruit Gram character tendency of the element 5 Blinding was not possible to be so different, surgery adjuvant therapy for patients with oral cancer resection for CR doxifluridine
Spanish	AI discharge five patients from group thiopental had died and eight in the group of pentobarbital- barbital (P0, 16). At six months remained the super- EXPERIENCE. Patients with affeccio- of the skin that could prevent serious evaluation therapy

Comparison of Extractions From Translated and Original Articles

Table 7 displays the adjusted percentage of correct extractions per language, including English, and per analyzed extraction item; the percentages are adjusted for individuals' likelihood of correctly extracting English articles. To recap, the reference standards for English-language articles were the consensus extraction across all researchers or between the senior investigators; the reference standards for translated articles were the double data extraction results from original language extractions. The extraction items are clustered by study domain (study design, intervention description data including the number of participants randomized,

outcome descriptions, and results). The specific extraction questions are described in Appendix A. In general, across languages the agreement between the extractors and the reference standards for each article (from consensus in English and from double-extracted original language articles in other languages) was greater for design and intervention domain items than for outcome descriptions and study results. In particular, extractors did relatively poorly extracting which outcomes from a given list were reported in the study and in extracting net differences (or equivalent results) and their standard errors for continuous outcomes.

Table 7. Percentage of correct extractions, per item and language, adjusted for individual's likelihood of correctly extracting the same data item from English articles

Domain	Extraction Item	English	Chinese	French	German	Japanese	Spanish
Design	Inclusion criteria	99	99	100	98	93	99
Design	Exclusion criteria	98	97	100	70	79	97
Design	Funding source	93	100	86	99	99	35
Design	No. centers	86	97	85	87	51	87
Design	Followup duration	88	79	41	82	84	96
Design	Randomization technique	95	87	82	78	89	93
Design	Allocation concealment method	93*	100*	85*	100*	80*	85*
Design	Intention-to-treat analysis	98*	100*	90*	95*	80*	75*
Design	Power calculation	99	95	74	97	100	98
Design	Subject blinding (explicit)	93	100	98	97	90	78
Design	Caregiver blinding (explicit)	92	100	95	96	98	98
Design	Outcome assessor blinding (explicit)	89	88	91	97	96	90
Design	Double blinded	94*	100*	100*	95*	100*	95*
Design	Single blinded	99*	100*	95*	100*	90*	100*
Design	Outcomes reported†	63	36	43	12	5	21
Design	No outcomes missed‡	75	80	74	13	11	31
Design	No extra outcomes added**	96	90	82	99	96	99
Intervention	Dose (of all interventions)	82	88	97	79	91	81
Intervention	Frequency (of all interventions)	87	74	87	86	75	83
Intervention	Route (of all interventions)	100	99	100	100	100	100
Intervention	Duration (of all interventions)	96	28	75	66	88	77
Intervention	No. randomized (for all interventions)	71	97	91	97	82	89
Outcome	Description	87	45	51	42	75	98
Results	No. analyzed (per intervention)	100	96	98	98	94	98
Results	Mean or median reported††	94*	100*	100*	84*	71*	100*
Results	Net difference††	81	28	15	73	67	60
Results	Standard error of net difference††	81	38	15	64	56	71
Results	No. events (counts) or odds ratio‡‡	99	98	86	37	47	94
Results	Reported P value of difference or odds ratio	93	30	81	73	21	73

Table 7. Percentage of correct extractions, per item and language, adjusted for individual's likelihood of correctly extracting the same data item from English articles (continued)

Domain	Extraction Item	English	Chinese	French	German	Japanese	Spanish
Overall***	% items ≥98% correct	30	41	26	22	19	30
Overall	% items ≥91% correct	63	59	37	48	30	48
Overall	% items ≥76% correct	93	74	74	70	67	78
Overall	% items ≥51% correct	100	78	85	89	89	93
Overall	% items ≤50% correct	0	22	15	11	11	7

Shading of cells matches reported percentages

98-100% correct 100th percentile)	91-97% correct 72nd percentile)	76-90% correct 50th percentile)	51-75% correct 24th percentile)	≤50% correct 11th percentile).

* Crude (unadjusted) percentage.
† From a list of proffered outcomes, there was exact agreement as to which were reported in the study.
‡ No outcomes found in the original article were missed from the translated article. This item is excluded from the overall percentages of items correct at the bottom of the table.
** No outcomes not found in the original article were added from the translated article. This item is excluded from the overall percentages of items correct at the bottom of the table.
†† For continuous outcomes.
‡‡ For dichotomous outcomes.
*** The five "Overall" rows display the percentage of the 29 items (not the individual extractions), per language, that were each extracted correctly the given percentage of the time (e.g., 30% of the English items were extracted correctly 98-100% of the time). Note that the final two rows, by definition, sum to 100%.

Translated Chinese articles yielded the largest percentage of items (22 percent) incorrectly extracted by more than half the extractors, although Chinese articles also yielded the largest percentage of items (41 percent) extracted correctly by more than 98 percent of the extractors (including English article extractions). However, translated Chinese articles had particularly lower likelihoods of correct extractions for the important extraction items about descriptions of the interventions, the outcomes, and the results. In contrast, extractors of translated Spanish articles had relatively high likelihoods of extracting items correctly except, in comparison with English, for results data. For Spanish, only 7 percent of items had less than 50 percent correct extractions, including funding source and identifying reported outcomes. Extractions of other translated language articles yielded similar patterns as for translated Chinese articles, but with generally higher rates of correct extractions. In particular, identifying reported outcomes and extracting results were more likely to be incorrect.

Table 8 displays the adjusted odds ratios between translated and English articles of correctly extracting individual items (the odds of correct extractions from translated articles versus the odds of correct extractions from English articles, adjusted for each researcher's likelihood of correctly extracting the English data items). Of note, it was not uncommon that the odds of correctly extracting individual items from the translated articles were greater than the odds of doing so from the 10 extracted English articles. All odds ratios of 1 or greater were analyzed as being equivalent to perfect agreement. Overall, the pattern of odds ratios of adjusted odds ratios of correct answers compared with English across items and languages (Table 8) was similar to the pattern of adjusted percentages (Table 7). It highlights that for all translated languages except Spanish, extractors were statistically significantly more likely to extract incorrect data for outcome description and results from translated articles than from English articles. Similarly, the likelihood of missing reported outcomes was higher from translated articles, significantly so for German, Japanese, and Spanish articles. The seeming discrepancy between Tables 7 and 8 in the results for duration of interventions (with the "Intervention" domain) is due to the near 100

percent accuracy for all languages and thus small numbers of incorrect extractions (e.g., 2 or 4 percent versus 0 percent).

Table 8. Odds ratios (confidence intervals) compared with English of correct extractions, adjusted for individual's likelihood of correctly extracting the same data item from English articles

Domain	Extraction Item	Chinese	French	German	Japanese	Spanish
Design	Inclusion criteria	1.9 (0.1, >5)	3.0 (0.2, >5)	0.5 (<0.1, >5)	0.1 (<0.1, 1.4)	0.8 (0.1, >5)
Design	Exclusion criteria	0.6 (<0.1, >5)	>5 (<0.1, >5)	<0.1 (<0.1, 0.7)	<0.1 (<0.1, 1.2)	0.7 (<0.1, >5)
Design	Funding source	>5 (0.2, >5)	0.5 (<0.1, >5)	>5 (0.2, >5)	>5 (0.2, >5)	<0.1 (<0.1, 1.0)
Design	No. centers	4.7 (0.4, >5)	0.9 (0.2, >5)	1.1 (0.2, >5)	0.2 (<0.1, 0.9)	1.1 (0.2, >5)
Design	Followup duration	0.5 (0.1, >5)	<0.1 (<0.1, 1.0)	0.7 (0.1, >5)	0.7 (0.1, >5)	3.8 (0.3, >5)
Design	Randomization technique	0.3 (<0.1, 2.2)	0.2 (<0.1, 1.2)	0.2 (<0.1, 1.0)	0.4 (0.1, 2.8)	0.7 (0.1, >5)
Design	Allocation concealment method	3.6 (0.2, >5)*	0.5 (0.1, 2.0)*	3.6 (0.2, >5)*	0.3 (0.1, 1.3)*	0.5 (0.1, 2.0)*
Design	Intention-to-treat analysis	1.3 (0.1, >5)*	0.2 (<0.1, 1.7)*	0.5 (<0.1, >5)*	0.1 (<0.1, 0.6)*	0.1 (<0.1, 0.4)*
Design	Power calculation	0.3 (<0.1, 3.9)	<0.1 (<0.1, 0.7)	0.4 (<0.1, >5)	>5 (<0.1, >5)	0.6 (<0.1, >5)
Design	Subject blinding (explicit)	>5 (<0.1, >5)	4.4 (0.2, 104.1)	2.5 (0.1, >5)	0.6 (<0.1, >5)	0.3 (<0.1, 4.1)
Design	Caregiver blinding (explicit)	>5 (<0.1, >5)	1.7 (0.2, >5)	2.0 (0.2, >5)	4.3 (0.2, >5)	3.3 (0.2, >5)
Design	Outcome assessor blinding	0.9 (0.2, >5)	1.2 (0.3, >5)	3.4 (0.4, >5)	2.8 (0.3, >5)	1.1 (0.2, >5)
Design	Double blinded	2.4 (0.2, >5)*	0.1 (0.2, >5)*	0.5 (0.1, >5)*	2.4 (0.2, >5)*	1.0 (0.1, >5)*
Design	Single blinded	>5 (<0.1, >5)*	1.9 (<0.1, 4)*	1.2 (<0.1, >0.5)*	0.8 (<0.1, 1.3)*	0.6 (<0.1, >5)*
Design	Outcomes reported†	0.3 (0.1, 1.8)	0.4 (0.1, 2.5)	<0.1 (<0.1, 0.5)	<0.1 (<0.1, 0.4)	0.2 (<0.1, 1.0)
Design	No outcomes missed‡	1.3 (0.2, >5)	0.9 (0.2, 4.7)	<0.1 (<0.1, 0.3)	<0.1 (<0.1, 0.4)	0.1 (<0.1, 0.8)
Design	No extra outcomes added**	0.4 (<0.1, >5)	0.2 (<0.1, 4.8)	3.0 (0.1, 5)	1.1 (<0.1, >5)	4.1 (0.1, >5)

Table 8. Odds ratios (confidence intervals) compared with English of correct extractions, adjusted for individual's likelihood of correctly extracting the same data item from English articles (continued)

Domain	Extraction Item	Chinese	French	German	Japanese	Spanish
Intervention	Dose (of all interventions)	1.6 (0.4, >5)	>5 (0.9, >5)	0.8 (0.2, 2.8)	2.1 (0.4, >5)	0.9 (0.2, 4.2)
Intervention	Frequency (of all interventions)	0.4 (<0.1, 4.0)	1.0 (0.1, >5)	0.9 (0.1, >5)	0.4 (0.1, 3.8)	0.7 (0.1, >5)
Intervention	Route (of all interventions)	<0.1 (<0.1, 2.9)	0.3 (<0.1, >5)	0.2 (<0.1, >5)	0.5 (<0.1, >5)	>5 (<0.1, >5)
Intervention	Duration (of all interventions)	<0.1 (<0.1, 0.2)	0.1 (<0.1, 0.9)	<0.1 (<0.1, 0.7)	0.3 (<0.1, 2.3)	0.1 (<0.1, 0.9)
Intervention	No. randomized (for all interventions)	>5 (1.1, >5)	4.3 (0.6, >5)	>5 (1.1, >5)	1.8 (0.3, >5)	3.2 (0.5, >5)
Outcome	Description	0.1 (<0.1, 0.8)	0.2 (<0.1, 1.5)	0.1 (<0.1, 1.0)	0.5 (0.1, 2.8)	>5 (0.6, >5)
Results	No. analyzed (per intervention)	<0.1 (<0.1, 0.4)	<0.1 (<0.1, 0.9)	<0.1 (<0.1, 0.9)	<0.1 (<0.1, 0.3)	<0.1 (<0.1, 1.3)
Results	Mean or median reported††	2.8 (0.2, >5)*	2.3 (0.1, >5)*	0.4 (0.1, 1.6)*	0.2 (<0.1, 0.7)*	1.8 (0.1, >5)*
Results	Net difference††	<0.1 (<0.1, 0.7)	<0.1 (<0.1, 0.4)	0.6 (0.1, 3.7)	0.5 (0.1, 3.6)	0.3 (<0.1, 3.2)
Results	Standard error of net difference††	0.1 (<0.1, 0.9)	<0.1 (<0.1, 0.5)	0.4 (0.1, 2.6)	0.3 (<0.1, 2.3)	0.6 (0.1, >5)
Results	No. events (counts) or odds ratio‡‡	0.8 (<0.1, >5)	<0.1 (<0.1, 1.9)	<0.1 (<0.1, 0.3)	<0.1 (<0.1, 0.5)	0.2 (<0.1, >5)
Results	Reported P value of difference or odds ratio	<0.1 (<0.1, 0.2)	0.3 (0.1, 2.0)	0.2 (<0.1, 1.1)	<0.1 (<0.1, 0.2)	0.2 (<0.1, 1.3)

Shading of cells matches reported percentages

Odds ratio (OR) ≥1	0.5<OR<1 (nonsignificant [NS])	0.1<OR≤0.5 (NS)	OR ≤ 0.1 (NS)	Statistically significant (OR<0.1).

* OR based on crude (unadjusted) proportions correct.
† From a list of proffered outcomes, there was exact agreement as to which were reported in the study.
‡ No outcomes found in the original article were missed from the translated article. This item is excluded from the overall percentages of items correct at the bottom of the table.
** No outcomes not found in the original article were added from the translated article. This item is excluded from the overall percentages of items correct at the bottom of the table.
†† For continuous outcomes.
‡‡ For dichotomous outcomes.

Risk of bias assessment typically had only slight agreement across languages and risk of bias questions. The median kappa across questions among the original language extractors (including for English articles) was 0.195 (full range -0.14, 0.78) and for extractors of translated articles was 0.22 (-0.46, 1.00). For English articles, 49 percent of biases were rated "unclear," 12 percent "high," and 39 percent "low." Among other original language articles, 42 percent of biases were

rated “unclear,” 18 percent “high,” and 39 percent “low.” Among translated articles, 53 percent of biases were rated “unclear,” 11 percent “high,” and 37 percent “low.” Table 9 displays the kappa values for each question, within each language, for both original and translated articles, along with P values for differences between original and translated extractions. Among 25 comparisons (5 questions in 5 languages), only 3 (12 percent) have a P value less than 0.10. Among Chinese and Spanish articles, allocation concealment was rated more consistently among translated than original articles (P = 0.06). This can be ascribed to the more universal designation of “unclear” bias among translated articles. Only for the designation of attrition bias among Chinese articles was agreement significantly poorer for translated articles (7 “unclear,” 1 “high,” 12 “low”) than for original articles (6 “unclear,” 3 “high,” 11 “low”).

Table 9. Kappa statistics for agreement of risk of bias

		English	Chinese	French	German	Japanese	Spanish
Random'n	Orig	0.37 (0.26, 0.48)	0.67 (0.21, 1.00)	0.33 (-0.07, 0.74)	0.62 (0.04, 1.00)	0.14 (-0.33, 0.61)	0.01 (-0.30, 0.33)
Method	Trx		0.63 (0.08, 1.00)	0.12 (-0.38, 0.61)	0.22 (-0.32, 0.76)	-0.31 (-0.85, 0.24)	0.39 (-0.23, 1.00)
	P		*0.91*	*0.50*	*0.33*	*0.22*	*0.28*
Allocation	Orig	0.28 (0.18, 0.38)	0.26 (-0.18, 0.70)	-0.14 (-0.57, 0.29)	0.78 (0.27, 1.00)	0.02 (-0.06, 0.10)	0.07 (-0.10, 0.24)
Conceal't	Trx		1.00 (0.38, 1.00)	-0.39 (-0.87, 0.10)	0.38 (-0.16, 0.92)	0.24 (-0.28, 0.76)	1.00 (0.38, 1.00)
Method	*P*		***0.06***	*0.46*	*0.29*	*0.42*	***<0.01***
Patient	Orig	0.18 (0.10, 0.27)	0.29 (-0.06, 0.63)	0.11 (-0.05, 0.28)	-0.00 (-0.43, 0.43)	0.26 (-0.12, 0.65)	0.50 (0.04, 0.96)
Caregiver	Trx		0.01 (-0.46, 0.48)	-0.08 (-0.55, 0.38)	0.21 (-0.24, 0.66)	0.07 (-0.37, 0.51)	0.34 (-0.13, 0.81)
Blinding	*P*		*0.35*	*0.44*	*0.51*	*0.51*	*0.63*
Outcome	Orig	0.28 (0.18, 0.37)	-0.11 (-0.55, 0.33)	0.10 (-0.12, 0.33)	0.29 (-0.05, 0.62)	-0.10 (-0.37, 0.18)	0.13 (-0.24, 0.50)
Assessor	Trx		-0.30 (-0.79, 0.18)	0.03 (-0.43, 0.49)	0.22 (-0.32, 0.76)	-0.46 (-0.93, 0.01)	0.51 (0.04, 0.98)
Blinding	*P*		*0.56*	*0.78*	*0.85*	*0.18*	*0.21*
Attrition	Orig	0.04 (-0.05, 0.13)	0.38 (0.04, 0.73)	0.00 (-1.00, -1.00)	0.21 (-0.18, 0.60)	0.29 (-0.16, 0.73)	0.05 (-0.15, 0.26)
Bias	Trx		-0.36 (-0.90, 0.18)	0.34 (-0.13, 0.81)	0.55 (0.02, 1.00)	0.45 (-0.10, 0.99)	0.31 (-0.19, 0.81)
	P		***0.02***	*	*0.31*	*0.65*	*0.35*

Formatting coding

κ<0 (poor agreement)	κ 0-0.20 (slight agreement)	κ 0.21-0.40 (fair agreement)	κ 0.41-0.60 (moderate agreement)	κ 0.61-0.80 (substantial agreement)	κ 0.81-1.00 (almost perfect agreement).

Bolded P values are those where P<0.10.

Conceal't = concealment, Orig = original language articles, P = P value (between original and translated articles), Random'n = randomization, Trx = translated articles.

* P value could not be estimated.

We performed sensitivity analyses where we dichotomized the risk of bias assessment by setting "unclear" to be equivalent to either "high" or "low" risk of bias. The only finding that was different between the main and sensitivity analyses was when "unclear" was set to be equivalent to "high" risk of bias, among translated Spanish articles, outcome assessor blinding was more consistently graded "high/unclear" (75 percent) than among original articles (60 percent), P = 0.06.

Association Between Extractor Confidence and Accuracy

Table 5 above displays the distribution of levels of confidence extractors had in the accuracy and completeness of the translations across languages. To examine the association between their level of confidence and their extraction accuracy, we first calculated the raw percentage accuracy by confidence level, by language and across languages (Table 10). For French articles the accuracy was considerably higher when extractors had strong confidence (94 percent accuracy across articles and items) than moderate or little confidence (67 percent accuracy). However, this pattern was not seen for other languages and of note, for Chinese articles the accuracy was higher when extractors had little confidence (88 percent) than moderate or strong accuracy (73 percent). Overall across all languages, the accuracy was about the same regardless of extractors' confidence level (76 or 79 percent).

Table 10. Association between extractors' confidence in accuracy of translation and their extraction accuracy

Confidence	Chinese, Percent Accurate	French,* Percent Accurate	German, Percent Accurate	Japanese, Percent Accurate	Spanish, Percent Accurate	Overall,* Percent Accurate
Strong	73%	94%	78%	66%	80%	79%
Moderate	78%	76%	77%	69%	83%	76%
Little	88%	67%	74%	75%	74%	76%

* 1 extractor did not rate confidence level for 1 article.

Discussion

Our results showed that using Google Translate to translate medical articles in many cases may be feasible and not a resource-intensive process that leads to operationally workable English versions. The accuracy of translation was heavily dependent on the original language of the article. Specifically, extractions of Spanish articles were most accurate, followed by fairly accurate extractions from German, Japanese, and French articles. The least accurate data extractions resulted from translated Chinese articles. With the exception of Japanese (where we found that extraction was fairly accurate) difference across languages was similar to the findings of machine translation experts for general translation "that translations between European languages are usually good, while those involving Asian languages are often relatively poor."[13]

With the exception of Spanish, the findings of this analysis are generally consistent with, but more robust than, a similar analysis done as a pilot study.[15] Our improved methods, including double data extraction of the original language articles together with adjustment for individual extractors' accuracy in extracting English articles provides better confidence in our conclusions. The discrepancy in the results from the translated Spanish articles are likely due to greater disagreement in data extraction (unrelated to translation issues) between individual pairs of extractors than between double data extracted and reconciled extractions and the translated extractions.

Across languages, including English, we found good levels of agreement (mostly above 85 to 90 percent) for extraction of most study design questions (eligibility criteria; funding source; number of centers; followup duration; whether the study reported randomization, allocation concealment techniques, intention-to-treat and power calculations; and who was blinded. With slightly lower agreement, there were also generally good levels of agreement (mostly above 70 percent) for extraction of descriptions of the intervention (dose, frequency, route, duration) and the number of participants randomized. For results reporting, there was consistently accurate extraction (mostly above about 85 percent) for the numbers of participants analyzed and whether mean or median data were reported. The odds ratios of accurate extractions compared with English followed similar patterns. The accuracy of descriptions of outcomes and results data (net difference, standard error, number of events or reported odds ratio, and P value of difference or odds ratio) varied widely by language, with descriptions of outcomes being commonly inaccurate from Chinese, German, and French articles, continuous results data (net difference and standard error) being inaccurate from French and Chinese articles, categorical results data (number of events or odds ratio) being inaccurate from German and Japanese articles, and P values being inaccurate in Japanese and Chinese articles. Translated Spanish articles generally yielded more accurate outcome descriptions and results data. Extractors' accuracy in finding reported outcomes from lists of outcomes was generally poor, including from English articles (with only 63 percent accuracy). Only 5 to 43 percent of translated articles yielded accurate lists of reported outcomes across languages. Most of the inaccuracy came from missing outcomes from the list, but a few arose from finding outcomes not captured in the reference extraction.

We expected to find that investigators would provide more accurate extractions when they had greater confidence in the accuracy and completeness of the translations. However, with the possible exception of French studies, we did not find this to be the case. It is unclear why the data extractors failed to be more confident about studies they more accurately extracted. It may be that they were unable to disambiguate difficulties in extracting the studies due to poor translation from those due to poor reporting. This finding should not be overinterpreted but it

does call into question whether extractors can subjectively assess how accurate their extractions from translated articles are.

Although our double data extraction of original language articles and the adjustment for accuracy of extraction of English language articles improved on the limitations of the pilot study, these approaches still do not fully remove the possibility that differences (or lack of differences) between languages that we found were in part due to intrinsic differences between data extractors or the different articles in the different languages. As we describe in the Methods section, we changed our approach to reconciliation of the reference standards to allow for multiple correct answers. We did this to reduce the number of disagreements that occurred between translated and original articles that were due to differences in interpretation of the data rather than translation errors. However, it remains likely that a number of the disagreements were due to differences in interpretation. Similarly, while we controlled for extractors' likelihood of extracting English articles correctly, we could not fully control for the likelihood that extractors made errors unrelated to translation in specific articles. While extractors each extracted articles from different languages, we did not achieve an even distribution of the extractors across languages. Furthermore, there were fundamental differences in the studies across different languages, in the medical fields being examined and the complexity of the study designs, interventions, outcomes, and analyses. These intrinsic differences may have resulted in some of the differences in accuracy of extraction. We did not have extractors of translated articles reconcile their extractions and then compare the reconciled translated and reconciled original language extractions. Doing so might have more closely mimicked typical systematic review methods, but would have greatly reduced the study's power. However, despite our power calculation, the confidence intervals of the adjusted odds ratios between translated and English articles were generally wide, possibly resulting in either an overestimation of the number of items with "trends" toward large differences in accuracy (i.e., small but nonsignificant odds ratios) or an underestimation of the number of true effects (due to frequent nonsignificance).

Other limitations that were described for the pilot study still hold. While native speakers were chosen to extract the original language articles, these extractors were not always medically trained in their native language. Thus, translations that employed non-English medical terminology may have been difficult to extract from the original articles. However, this limitation should have been mitigated by the double data extraction. All extractors may or may not have been familiar with the medical topic covered by the article, which is another factor introducing variability to the results. It is likely that the data extraction error rate was higher than for a typical systematic review, since the articles were on random topics and the data extractors were neither trained nor necessarily proficient in the clinical domains.

The Google Translate tool is ever evolving and presumably improving, as users around the world improve the accuracy of translations. It is also reasonable to assume that with time more articles from more non-English language publications will be in a format that can be directly (and thus quickly) translated. However, this also implies that the accuracy of translations between different pairs of languages will at least partly depend on how many words and documents are being translated among different languages on the Internet. While we did not test for differences based on different study countries, it is of interest to note that half the Spanish articles were written in Spain and half in Latin America and half the French articles were written in France and half distributed among Canada (Quebec), Tunisia, and Turkey. All the Chinese articles were written in China in simplified Chinese. Anecdotally, it is our experience that extremely few studies from other Chinese-speaking countries or territories (Taiwan, Singapore,

Hong Kong, Macao) are written in Chinese, particularly in traditional Chinese. Our analysis is relevant only for simplified Chinese. Although data extraction from translated articles was assessed to be considerably more difficult and time consuming than extraction from equivalent English language articles, extraction was always feasible in what was considered to be a reasonable amount of time, even including the extra time required to perform article translation. For this research project, we used the directly available Google Translate Web site (translate.google.com) not the Google Translate Toolkit (translate.google.com/toolkit/), which requires an account setup and login. For typical systematic reviews, the toolkit may offer some advantages including the feature that it searches for previous human translations of the same text and allows improved translations on the fly. However, this feature would likely be of value only if the investigator himself or herself, as opposed to a research assistant, does the translation and puts in the effort to critically evaluate the translation.

Even though Google translation of medical articles in most cases is far from perfect and on average results in higher levels of inaccuracies than extraction from English, we conclude that the technique has potential to be of value and that for most of the tested languages it may be reasonable to attempt translation (with Google Translate) and extraction of non–English-language articles that are available as machine-readable PDF (or HTML) files. A major caveat, though, is that we found that extraction of results data were least accurate. Thus, extra care should be taken when considering how much to rely on or accept the results data from translated articles. It would be appropriate to consistently perform sensitivity analyses regarding translated articles, where possible differences in findings (by meta-analysis) or conclusions (overall) may occur when translated articles are included or omitted. It should be recognized that any differences may be due not only to differences in applicability or methodology, but to errors in translation. Our prior anecdotal experience suggests that using Google Translate for articles in languages that an extractor is at least somewhat familiar with can be particularly useful to allow confident data extraction. Based on the evidence that machine translation is (only) mostly accurate and our anecdotal experience, an appropriate approach for systematic reviewers may be to run the machine translation and have a native speaker confirm or revise the translations. If such human translators are available, this may be a time- and cost-efficient approach. A reasonable alternative conclusion, however, is that the translation software is still sufficiently inaccurate for use in systematic review, that the risk of introducing errors is too great. Each investigator considering the inclusion of articles requiring machine translation into a systematic review will need to decide the appropriate balance between completeness and risk of extraction errors.

The value and reliability of machine translation of articles for systematic review requires further research. Questions of interest include: Are the findings of this study replicable with a different set of articles and extractors (we would suggest that if feasible, a larger sample of studies be tested)? How do different machine translators compare? How does machine translation from other languages fare (although, the value of testing languages with relatively few publications is limited)? Are there differences in extraction accuracy based on differences in study designs, including differences in clinical or content areas, pharmacological versus nonpharmacological interventions, different outcome types, or randomized versus nonrandomized studies? How would the data extraction errors from poor translation impact meta-analysis results and systematic review conclusions?

We conclude that it is reasonable for systematic reviewers to devote the small amount of resources and effort necessary to try Google Translate to include non-English articles. It will be

important, however, to recognize that extraction of these articles is more prone to error than extraction of typical English language articles. Therefore, judgment will be needed to determine how much confidence the reviewers have in the accuracy of the data extraction of these articles, and to recognize that apparently missing data or unclearly reported data may be more a factor of poor translation than of poor methodology.

References

1. Egger M, Zellweger-Zahner T, Schneider M, et al. Language bias in randomised controlled trials published in English and German. Lancet. 1997;350(9074):326-9. PMID 9251637.
2. Committee on Standards for Systematic Reviews of Comparative Effectiveness Research. Finding what works in health care: standards for systematic reviews. Washington, DC. 2011.
3. Heres S, Wagenpfeil S, Hamann J, et al. Language bias in neuroscience—is the Tower of Babel located in Germany? Eur Psychiatry. 2004;19(4):230-2. PMID 15196606.
4. Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess. 2003;7(1):1-76. PMID 12583822.
5. Gregoire G, Derderian F, Le LJ. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol. 1995;48(1):159-63. PMID 7853041.
6. Juni P, Holenstein F, Sterne J, et al. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002;31(1):115-23. PMID 11914306.
7. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol. 2000;53(9):964-72. PMID 11004423.
8. Morrison A, Moulton K, Clark M , et al. English-language restriction when conducting systematic review-based meta-analyses: systematic review of published studies. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2009. http://cadth.ca/media/pdf/H0478_Language_Restriction_Systematic_Review_Pub_Studies_e.pdf. Accessed August 30, 2012.
9. Pham B, Klassen TP, Lawson ML, et al. Language of publication restrictions in systematic reviews gave different results depending on whether the intervention was conventional or complementary. J Clin Epidemiol. 2005;58(8):769-76. PMID 16086467.
10. Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408-12. PMID 7823387.
11. Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. Int J Epidemiol. 2002;31(1):150-3. PMID 11914311.
12. Aiken M, Ghosh K, Wee J , et al. An evaluation of the accuracy of online translation systems. Communications of the IIMA (International Information Management Association). December 2009. http://findarticles.com/p/articles/mi_7099/is_4_9/ai_n56337599/?tag=content. Accessed 11 Jul 2012.
13. Aiken M, Balan S. An analysis of Google Translate accuracy. Translation Journal. 16[2]. April 2011. http://translationjournal.net/journal/56google.htm. Accessed 10 Jul 2012
14. Freitas de Souza R, Sequeira P, Nasser M , et al. [P08-5] Is Google Translate useful for the selection of studies to be included in Cochrane reviews? 17th Cochrane Colloquium, Singapore, 11-14 October 2009. http://www.imbi.uni-freiburg.de/OJS/cca/index.php?journal=cca&page=article&op=view&path%5B%5D=8037. Accessed 11 Jul 2012.
15. Balk EM, Chung M, Hadar N, et al. Accuracy of Data Extraction of Non-English Language Trials With Google Translate. Methods Research Report. (Prepared by the Tufts Evidence-based Practice Center under Contract No. 290-2007-10055 I.) AHRQ Publication No. 12-EHC056-EF. Rockville, MD: Agency for Healthcare Research and Quality. April 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

16. Ip S, Hadar N, Keefe S, et al. A Web-based archive of systematic review data. Systematic Reviews. 2012;1:15. http://www.systematicreviewsjournal.com/content/1/1/15.

17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-74. PMID 843571.

Appendix A. Data Extraction Items

Domain	Extraction Item	Entry type	Options	Analyzed?
Design	Inclusion criteria	Free text	--	Yes
Design	Exclusion criteria	Free text	--	Yes
Design	Funding source	Check boxes with free text option (multiple choice allowed)	• Government • Industry • Academic or Hospital • Foundation etc. • Named entity of an unclear category (free text for name) • No funding (explicitly stated) • Not reported	Yes
Design	No. centers	Radio buttons with free text option (single answer allowed)	• Single center • Multicenter (how many centers? • Not reported/unclear	Yes
Design	Followup duration (maximum or mean)	Free text	--	Yes
Design	Randomization technique reported?	Radio buttons	• Yes • No	Yes
Design	Allocation concealment method reported?	Radio buttons	• Yes • No	Yes
Design	Intention-to-treat analysis reported for any outcome (either in methods or results sections)	Radio buttons	• Yes • No	Yes
Design	Power calculation reported for any outcome	Radio buttons	• Yes • No	Yes
Design	Subject blinding explicitly reported (for any outcome)	Radio buttons	• Yes • No	Yes
Design	Caregiver blinding explicitly reported (for any outcome)	Radio buttons	• Yes • No	Yes
Design	Outcome assessor blinding explicitly reported (for any outcome)	Radio buttons	• Yes • No	Yes
Design	Double blinded(for any outcome)	Radio buttons	• Yes • No	Yes

Design	Single blinded(for any outcome)	Radio buttons	• Yes • No	Yes
Design	Select all outcomes reported from a given list of about 12 outcomes	Check boxes (multiple choices allowed)	• Different outcomes for each language culled from the articles • None of these outcomes found	Yes
Intervention	Study arm title	Free text	--	No
Intervention	Study arm description	Free text	--	No (inadequate nonredundant data were extracted)
Intervention	Dose (individually for each outcome)	Free text	--	Yes (for 1 or 2 preselected outcomes only)
Intervention	Frequency (individually for each outcome)	Free text	--	Yes (for 1 or 2 preselected outcomes only)
Intervention	Route (individually for each outcome)	Free text	--	Yes (for 1 or 2 preselected outcomes only)
Intervention	Duration (individually for each outcome)	Free text	--	Yes (for 1 or 2 preselected outcomes only)
Intervention	No. randomized into each study arm (for all interventions)	Free text	--	Yes (for 1 or 2 preselected outcomes only)
Outcome	Outcome title	Free text	--	No (1 preselected continuous and/or 1 preselected categorical outcome)
Outcome	Outcome units	Free text	--	No (inadequate meaningful data were extracted)
Outcome	Outcome description	Free text	--	Yes
Results	Continuous: No. analyzed (per intervention)	Free text	--	Yes
Results	Continuous: Mean or median reported††	Free text	“Mean” or “Median” (not the value)	Yes
Results	Continuous: Net difference (or difference between final)	Free text	(calculate if necessary)	Yes
Results	Continuous: Standard error of net difference	Free text	(calculate if necessary)	Yes
Results	Categorical: No. analyzed (per intervention)	Free text	--	Yes
Results	Categorical: No. events (counts)	Free text	--	Yes (combined with reported odds ratio and confidence interval)
Results	Categorical: Odds ratio and 95% confidence interval	Free text	(calculate if necessary)	Only if there were no counts data (only reported data; calculations not analyzed)

Results	Both: Reported P value of difference or odds ratio	Free text	--	Yes
Results	Both: What analysis adjusted for	Free text	--	No (insufficient data)
Study Quality	What is the risk of selection bias (biased allocation to interventions) due to inadequate generation of a randomized sequence?	Pull down menu	• Low • Unclear • High	Yes
Study Quality	What is the risk of selection bias (biased allocation to interventions) due to inadequate concealment of allocations before assignment?	Pull down menu	• Low • Unclear • High	Yes
Study Quality	For the preselected categorical outcome (or continuous outcome if there is no categorical outcome), what is the risk of performance bias due to knowledge of the allocated interventions by participants and personnel during the study (lack of study participant and personnel blinding)?	Pull down menu	• Low • Unclear • High	Yes
Study Quality	For the preselected categorical outcome (or continuous outcome if there is no categorical outcome), what is the risk of detection bias due to knowledge of the allocated interventions by outcome assessment (lack of outcome assessor blinding)?	Pull down menu	• Low • Unclear • High	Yes
Study Quality	For the preselected categorical outcome (or continuous outcome if there is no categorical outcome), what is the risk of attrition bias due to amount, nature, or handling of incomplete outcome data?	Pull down menu	• Low • Unclear • High	Yes
Miscellaneous	For TRANSLATED articles: How much additional time do you estimate the extraction of the translated article took compared to an extraction of a similar English article?	Free text	• <5 minutes extra time • 6-30 min extra time • >30 min extra time	Yes
Miscellaneous	For TRANSLATED articles: How confident are you in the accuracy and completeness of the translation of the original article?	Free text	• I have little confidence • I have a moderate confidence * • I have strong confidence	Yes
Miscellaneous	For TRANSLATED articles: Please provide examples of poor translations that made extraction difficult	Free text	--	C

Appendix B. List of Translated and Included Articles

Chinese

Hu XY, Zhou YX, Xu SZ, Lin YY. [Effects of probiotics on feeding intolerance in low birth weight premature infants]. [Chinese]. Zhongguo Dang Dai Er Ke Za Zhi. 2010;12:693-5. PMID 20849715.

Li H, Dong L, Li Y, Fu S. [A randomized clinical trial of combination of Aidi injection with Gemcitabine and Oxaliplatin regimen or Go regimen only in the treatment of advanced non-small-cell lung cancer.]. [Chinese]. Zhongguo Fei Ai Za Zhi. 2008;11:570-3. PMID 20735973.

Liu X, Liu D, Li J, Ou D, Zhou Z. [Safety and efficacy of carbon dioxide insufflation during colonoscopy]. [Chinese]. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2009;34:825-9. PMID 19734597.

Tang FZ, Liu YL, Wen FQ, Zhang ZX. [Comparison of therapeutic effects in severe nocturia: gradual versus immediate drug withdrawal]. [Chinese]. Zhongguo Dang Dai Er Ke Za Zhi. 2010;12:198-200. PMID 20350430.

Wang P, Yang J, Liu G, Chen H, Yang F. [Effects of moxibustion at head-points on levels of somatostatin and arginine vasopressin from cerebrospinal fluid in patients with vascular dementia: a randomized controlled trial]. [Chinese]. Zhong Xi Yi Jie He Xue Bao. 2010;8:636-640. PMID 20619139.

Xu JS, Yang JW, Gu MN, Chen YM. [Effects of fentanyl on EC50 of ropivacaine for postoperative epidural analgesia after gynecological surgery]. [Chinese]. Di Yi Jun Yi Da Xue Xue Bao. 2004;24:1326-7. PMID 15567796.

Xu XH, Chang YT, Li L, Li J, Zhang DM, Zou XH. [Effect of fructose-1,6-diphosphete on myocardial preservation during pulmonary operations]. [Chinese]. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2008;33:966-9. PMID 19001742.

Yang MH, Li M, Dou YQ, et al. [Effects of Bushen Huoxue Granule on motor function in patients with Parkinson's disease: a multicenter, randomized, double-blind and placebo-controlled trial]. [Chinese]. Zhong Xi Yi Jie He Xue Bao. 2010;8:231-7. PMID 20226144.

Yi JH, Li RR. [Influence of near-work and outdoor activities on myopia progression in school children]. [Chinese]. Zhongguo Dang Dai Er Ke Za Zhi. 2011;13:32-5. PMID 21251384.

Zhang GQ, Ge L, Ding W, Li HJ. [The value of portal vein chemotherapy after radical resection in delaying intrahepatic recurrence of stage II primary hepatocellular carcinoma]. [Chinese]. Ai Zheng. 2008;27:1297-301. PMID 19079997.

French

Aubin M, Vezina L, Maziade J, Robitaille NM. [Control of arterial hypertension: effectiveness of an intervention performed by family practitioners]. [French]. Can Fam Physician. 1994;40:1742-52. PMID 7950469.

Aydin A, Karadayi K, Aykan U, Can G, Colakoglu K, Bilge AH. [Effectiveness of topical ciclosporin A treatment after excision of primary pterygium and limbal conjunctival autograft]. [French]. J Fr Ophtalmol. 2008;31:699-704. PMID 18971855.

Baillargeon L, Drouin J, Desjardins L, Leroux D, Audet D. [The effects of Arnica Montana on blood coagulation. Randomized controlled trial]. [French]. Can Fam Physician. 1993;39:2362-7. PMID 7903572.

Devogelaere T, Beresniak A, Raymaeckers A, Naacke H, Ssi YK, I, Bremond-Gignac D. [Clinical study of Supranettes pads in the treatment of seasonal or perennial allergic conjunctivitis in children]. [French]. J Fr Ophtalmol. 2006;29:593-8. PMID 16885888.

Fekih M, Ben ZN, Jnifen A, et al. [Comparing two Prepidil gel regimens for cervical ripening before induction of labor at term: a randomized trial]. [French]. J Gynecol Obstet Biol Reprod (Paris). 2009;38:335-40. PMID 19467806.

Gadioux-Madern F, Lelez ML, Sellami L, et al. [Influence of the instillation of two versus three eyedrops of cyclopentolate 0.5% on refraction of Caucasian nonstrabismic children]. [French]. J Fr Ophtalmol. 2008;31:51-5. PMID 18401299.

Gosselin P, Verreault R, Gaudreault C, Guillemette J. [Dietary treatment of mild to moderate hypercholesterolemia. Effectiveness of different interventions]. [French]. Can Fam Physician. 1996;42:2160-7. PMID 8974552.

Lamouliatte H, Perie F, Joubert-Collin M. [Treatment of Helicobacter pylori infection with lansoprazole 30 mg or 60 mg combined with two antibiotics for duodenal ulcers]. [French]. Gastroenterol Clin Biol. 2000;24:495-500. PMID 10891736.

Polonovski JM, El MM. [Treatment of acute maxillary sinusitis in adults. Comparison of cefpodoxime-proxetil and amoxicillin-clavulanic acid]. [French]. Presse Med. 2006;35:33-8. PMID 16462661.

Rolachon A, Kezachian G, Causse X, et al. [Value of high-dose interferon-alpha in chronic viral hepatitis C patients non-responder to a 1st treatment. Pilot study prospective and randomized trial]. [French]. Gastroenterol Clin Biol. 1997;21:924-8. PMID 9587555.

German

Bechdolf A, Pohlmann B, Guttgemanns J, et al. [State-dependent motivational interviewing for people with schizophrenia and substance use : Results of a randomised controlled trial]. [German]. Nervenarzt. 2012;83:888-96. PMID 21720841.

Birnbaum F, Schwartzkopff J, Bohringer D, Reinhard T. [Penetrating keratoplasty with intrastromal corneal ring. A prospective randomized study]. [German]. Ophthalmologe. 2008;105:452-6. PMID 17899113.

Borner M, Burkle H, Trojan S, Horoshun G, Riewendt HD, Wappler F. [Intra-articular ketamine after arthroscopic knee surgery. Optimisation of postoperative analgesia]. [German]. Anaesthesist. 2007;56:1120-7. PMID 17726586.

Langer C, Forster H, Konietschke F, et al. [Mesh shrinkage in hernia surgery: data from a prospective randomized double-blinded clinical study]. [German]. Chirurg. 2010;81:735. PMID 20186380.

Marx S, Cimniak U, Beckert R, Schwerla F, Resch KL. [Chronic prostatitis/chronic pelvic pain syndrome. Influence of osteopathic treatment - a randomized controlled study]. [German]. Urologe A. 2009;48:1339-45. PMID 19705093.

Meybohm P, Hanss R, Bein B, et al. [Comparison of premedication regimes. A randomized, controlled trial]. [German]. Anaesthesist. 2007;56:890-6. PMID 17551699.

Schnabel M, Vassiliou T, Schmidt T, et al. [Results of early mobilisation of acute whiplash injuries]. [German]. Schmerz. 2002;16:15-21. PMID 11845337.

Stoffels I, Wolter TP, Sailer AM, Pallua N. [The impact of silicone spray on scar formation. A single-center placebo-controlled double-blind trial]. [German]. Hautarzt. 2010;61:332-8. PMID 19967328.

Warlo I, Krummenauer F, Dick HB. [Rotational stability in intraocular lenses with C-loop haptics versus Z haptics in cataract surgery. A prospective randomised comparison]. [German]. Ophthalmologe. 2005;102:987-92. PMID 15785909.

Wohlrab D, Droege JW, Mendel T, et al. [Minimally invasive vs. transgluteal total hip replacement. A 3-month follow-up of a prospective randomized clinical study]. [German]. Orthopade. 2008;37:1121-6. PMID 18810386.

Japanese

Adachi Y, Sumikuma T, Kagami R, et al. [Improvement of patient adherence by mixing oral itraconazole solution with a beverage (orange juice)]. [Japanese]. Rinsho Ketsueki. 2010;51:315-9. PMID 20534951.

Hirata K, Nakahara S, Shimokobe T, et al. [A randomized controlled trial of postoperative adjuvant chemotherapy for colorectal cancer-optimal duration of the treatment]. [Japanese]. Gan To Kagaku Ryoho. 2009;36:77-82. PMID 19151567.

Kurokawa M, Masuda Y, Noda M, et al. [Minimal effective dose on serum cholesterol concentration and the safety evaluation of dressing containing plant sterol in Japanese subjects]. [Japanese]. J Oleo Sci. 2008;57:23-33. PMID 18075220.

Miura H, Takahashi Y, Kitabatake Y. [Influence of group training on pulse wave velocity in elderly women]. [Japanese]. Nihon Koshu Eisei Zasshi. 2010;57:271-8. PMID 20560409.

Mochizuki M, Hatsugaya M, Rokujoh E, et al. [Randomized controlled study on the effectiveness of community pharmacists' advice for smoking cessation by Nicorette--evaluation at three months after initiation]. [Japanese]. Yakugaku Zasshi. 2004;124:989-95. PMID 15577269.

Satou Y, Kanda J, Okumura M, Nishida K. [An analysis of the educational effects of group counseling with visual aids: efforts to prevent diabetes in a business office setting]. [Japanese]. Sangyo Eiseigaku Zasshi. 2004;46:117-21. PMID 15382712.

Sawada A, Sakata N, Higuchi B, et al. [Comparison of micafungin and fosfluconazole as prophylaxis for invasive fungal infection during neutropenia in children undergoing chemotherapy and hematopoietic stem cell transplantation]. [Japanese]. Rinsho Ketsueki. 2009;50:1692-9. PMID 20068276.

Sekine Y, Takai Y, Nishii O, et al. [Establishment of an optimum bowel preparation method before gynecologic laparoscopic surgery]. [Japanese]. Yakugaku Zasshi. 2001;121:637-45. PMID 11523124.

Sugiura M, Hata Y, Fukuda T, et al. [One-week application of terbinafine cream compared with four-week application in treatment of Tinea pedis]. [Japanese]. Nihon Ishinkin Gakkai Zasshi. 2001;42:223-8. PMID 11704752.

Takahashi M, Araki A, Ito H. [Development of a new method for simple dietary education in elderly individuals with diabetes mellitus]. [Japanese]. Nihon Ronen Igakkai Zasshi. 2002;39:527-32. PMID 12404749.

Spanish

Bonetto G, Salvatico E, Varela N, Cometto C, Gomez PF, Calvo B. [Pain prevention in term neonates: randomized trial for three methods]. [Spanish]. Arch Argent Pediatr. 2008;106:392-6. PMID 19030637.

Ceriani Cernadas JM, Carroli G, Pellegrini L, et al. [The effect of early and delayed umbilical cord clamping on ferritin levels in term infants at six months of life: a randomized, controlled trial]. [Spanish]. Arch Argent Pediatr. 2010;108:201-8. PMID 20544134.

de Luis DA, de la FB, Izaola O, et al. [Randomized clinical trial with a inulin enriched cookie on risk cardiovascular factor in obese patients]. [Spanish]. Nutr Hosp. 2010;25:53-9. PMID 20204256.

Garcia-Talavera Espin NV, Gomez Sanchez MB, Zomeno Ros AI, et al. [Comparative study of two enteral feeding formulas in hospitalized elders: casein versus soybean protein]. [Spanish]. Nutr Hosp. 2010;25:606-12. PMID 20694297.

Gomez-Garcia A, Hernandez-Salazar E, Gonzalez-Ortiz M, Martinez-Abundis E. [Effect of oral zinc administration on insulin sensitivity, leptin and androgens in obese males]. [Spanish]. Rev Med Chil. 2006;134:279-84. PMID 16676098.

Lopez-De-Blanc SA, Salati-De-Mugnolo N, Femopase FL, et al. Antifungal topical therapy in oral chronic candidosis. A comparative study. Med Oral. 2002;7:260-70. PMID 12134127.

Martinez Gonzalez JM, Benito PB, Fernandez CF, San Hipolito ML, Penarrocha DM. A comparative study of direct mandibular nerve block and the Akinosi technique. Med Oral. 2003;8:143-9. PMID 12618675.

Perez-Barcena J, Barcelo B, Homar J, et al. [Comparison of the effectiveness of pentobarbital and thiopental in patients with refractory intracranial hypertension. Preliminary report of 20 patients]. [Spanish]. Neurocirugia (Astur). 2005;16:5-12. PMID 15756405.

Rodriguez MC, Castano SC, Garcia OL, Recio Rodriguez JI, Castano SY, Gomez Marcos MA. [Efficacy of an educational intervention group on changes in lifestyles in hypertensive patients in primary care: a randomized clinical trial]. [Spanish]. Rev Esp Salud Publica. 2009;83:441-52. PMID 19701575.

Vasquez AM, Sanin F, Alvarez LG, Tobon A, Rios A, Blair S. [Therapeutic efficacy of a regimen of artesunate-mefloquine-primaquine treatment for Plasmodium falciparum malaria and treatment effects on gametocytic development]. [Spanish]. Biomedica. 2009;29:307-19. PMID 20128355.

English

Davidson JA, Einhorn D, Allweiss P, et al. Effect of premixed nph and regular insulin on glucose control and health-related quality of life in patients with type 2 diabetes mellitus. Endocr Pract. 1997;3:331-6. PMID 15251769.

Friedman Z, Katznelson R, Phillips SR, et al. A randomized double-blind comparison of a morphine-fentanyl combination vs. morphine alone for patient-controlled analgesia following bowel surgery. Pain Pract. 2008;8:248-52. PMID 18503621.

Fu S, Choy NL, Nitz J. Controlling balance decline across the menopause using a balance-strategy training program: a randomized, controlled trial. Climacteric. 2009;12:165-76. PMID 19058060.

Halkes PH, van GJ, Kappelle LJ, Koudstaal PJ, Algra A. Medium intensity oral anticoagulants versus aspirin after cerebral ischaemia of arterial origin (ESPRIT): a randomised controlled trial. Lancet Neurol. 2007;6:115-24. PMID 17239798.

Ize-Iyamu IN, Saheeb BD. Feeding intervention in cleft lip and palate babies: a practical approach to feeding efficiency and weight gain. Int J Oral Maxillofac Surg. 2011;40:916-9. PMID 21641186.

Lyytinen J, Kaakkola S, Gordin A, Kultalahti ER, Teravainen H, Sovijarvi A. The effect of COMT inhibition with entacapone on cardiorespiratory responses to exercise in patients with Parkinson's disease. Parkinsonism Relat Disord. 2002;8:349-55. PMID 15177064.

Malpuech-Brugere C, Mouriot J, Boue-Vaysse C, et al. Differential impact of milk fatty acid profiles on cardiovascular risk biomarkers in healthy men and women. Eur J Clin Nutr. 2010;64:752-9. PMID 20485306.

Odergren A, Algvere PV, Seregard S, Libert C, Kvanta A. Vision-related function after low-dose transpupillary thermotherapy versus photodynamic therapy for neovascular age-related macular degeneration. Acta Ophthalmol. 2010;88:426-30. PMID 20597872.

Solon FS, Sarol JN, Jr., Bernardo AB, et al. Effect of a multiple-micronutrient-fortified fruit powder beverage on the nutrition status, physical fitness, and cognitive performance of schoolchildren in the Philippines. Food Nutr Bull. 2003;24:S129-S40. PMID 17016955.

Wang WC, Morales KH, Scher CD, et al. Effect of long-term transfusion on growth in children with sickle cell anemia: results of the STOP trial. J Pediatr. 2005;147:244-7. PMID 16126058.