

Accuracy of Data Extraction of Non-English Language Trials With Google Translate



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

Accuracy of Data Extraction of Non-English Language Trials With Google Translate

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 No. 290-2007-10055 I

Prepared by:

Tufts Evidence-based Practice Center,
Tufts Medical Center, Boston, MA

Investigators:

Ethan M. Balk, M.D., M.P.H.
Mei Chung, Ph.D., M.P.H.
Nira Hadar, M.S.
Kamal Patel, M.P.H., M.B.A.
Winifred W. Yu, Ph.D., R.D.
Thomas A. Trikalinos, M.D., Ph.D.
Lina Kong Win Chang, B.S.

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 was written with support from the Effective Health Care Program at AHRQ. None of the authors has a financial interest in any of the products discussed in this document. This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted, for which further reproduction is prohibited without the specific permission of copyright holders.

The investigators have no relevant financial interests in the report. The investigators have no employment, consultancies, honoraria, or stock ownership or options, or royalties from any organization or entity with a financial interest or financial conflict with the subject matter discussed in the report.

Suggested citation: Balk EM, Chung M, Hadar N, Patel K, Yu WW, Trikalinos TA, Chang L. 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.

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
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 acknowledge and thank the following highly experienced investigators who extracted non-English language articles. Without their input, this investigation would have been impossible: Johanna Paola Contreras, M.D. MSc, The Zena and Michael A. Wiener Cardiovascular Institute, Mount Sinai School of Medicine, New York, NY; Issa Dahabreh, M.D., M.S., Tufts Evidence-based Practice Center, Tufts Medical Center, Boston MA; Georgios D. Kitsios, M.D. Ph.D., Tufts Evidence-based Practice Center, Tufts Medical Center, Boston MA; Jounghee Lee, Ph.D., Graduate School of Education, Kyonggi University, Seoul, South Korea; Teruhiko Terasawa, M.D., Ph.D., Fujita Health University, School of Medicine, Tsu, Japan; Katrin Uhlig, M.D., M.S., Tufts Evidence-based Practice Center, Tufts Medical Center, Boston MA; and Linjie Zhang, M.D., Maternal and Child Health, Federal University of Rio Grande, Rio Grande do Sul, Brazil.

Accuracy of Data Extraction of Non-English Language Trials With Google Translate

Structured Abstract

Background: Systematic review prides itself on inclusion of all relevant evidence. However, study eligibility is often restricted to English language for practical reasons. Google Translate, a free Web-based resource for translation, has recently become available. However, it is unknown whether its translation accuracy is sufficient for Evidence-based Practice Center (EPC) systematic reviews. Therefore, we formally evaluated the accuracy of Google Translate for the purpose of data extraction of non-English language articles.

Methods: We retrieved 10 randomized controlled trials (RCTs) in eight languages (Chinese, French, German, Italian, Japanese, Korean, Portuguese, and Spanish) and eight observational studies in Hebrew. Eligible studies were RCTs that reported per-treatment group results data (except for Hebrew language studies, where no RCTs were identified). Each article was translated into English using Google Translate. The time required to translate each study was tracked. Data from the original language versions of the articles were extracted by one of 10 fluent speakers who were current or former members of our EPC. The English translated versions of the articles were extracted by one of five current EPC researchers who did not speak the given language. These five researchers also double data extracted 10 English language RCTs. Data extracted included: eligibility criteria, treatment description, study descriptors, quality issues, outcome description, and results. Extractors were also asked to estimate how much extra time was required for extraction compared to a similar English language article. For each study, pairs of data extractions were compared for agreement of each extracted item. We analyzed the percent agreement within sets of studies in each language for each extraction item and for groups of extraction items. We defined “high agreement” as at least 80 percent agreement within an item or article. The degree of agreement for each language was compared with that of the English language study comparisons with nonparametric tests.

Results: The length of time required to translate articles ranged from seconds (51 articles, 58 percent) to about 1 hour. Assessment by the English language data extractors 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 percent, 24 percent, 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.

Conclusion: Translation was not always possible, but generally required few resources. Across all languages, data extraction from translated articles was less accurate than from English language articles. Accurate extraction was possible for some articles in all languages, except

Chinese, with Portuguese and German articles yielding the most accurate extractions. 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.

Contents

Introduction	1
Aims	3
Methods	5
Study Selection	5
Translation	6
Basic Instructions Compiled for Article Translation	6
Data Extraction	6
Data Extraction Form	7
Data Extraction Comparison	7
Analysis.....	8
Results	10
Study Selection	10
Article Translation	10
Data Extraction From Translated Articles	11
Comparison of Translated With Original Articles.....	12
Discussion	23
References	27
Abbreviations	28

Tables

Table 1. Percentage of Studies From Medline in Various Languages.....	2
Table 2. Translation Time, by Language	10
Table 3. Agreement Across all Items, by Language	12
Table 4. Analyzed Data Extraction Items, by Section	13
Table 5. Percent of Items for Which There was Agreement, by Language and Data Extraction Form Section.....	14
Table 6. Items With at Least 80 Percent Agreement, by Language and Data Extraction Form Section	15

Figures

Figure 1. Histograms of the Percent Agreement for All Items, by Language	16
Figure 2. Histograms of the Percent Agreement for Eligibility Criteria Items, by Language	17
Figure 3. Histograms of the Percent Agreement for Descriptions of Treatment and Control Items, by Language	18
Figure 4. Histograms of the Percent Agreement for Study Characteristics Items, by Language	19
Figure 5. Histograms of the Percent Agreement for Study Methodology Items, by Language ..	20
Figure 6. Histograms of the Percent Agreement for Descriptions of Outcomes Items, by Language	21
Figure 7. Histograms of the Percent Agreement for Categorical Results Items, by Language ..	22
Figure 8. Histograms of the Percent Agreement for Continuous Results Items, by Language ..	23

Appendixes

- Appendix A. Annotated Data Extraction Forms
- Appendix B. List of Articles Translated and Included
- Appendix C. Examples of Poorly Translated Articles

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 recent Comparative Effectiveness Review (CER) reports 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 are at risk of selection bias based on language¹ 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 December 1, 2011 found that of 2,856,102 citations, 92 percent were published in English. Table 1 shows the number and frequency of publications in other languages with at least 1 citation.

In general, the Tufts EPC restricts to English language publications for its CERs and other reviews for which we expect large volumes of evidence. Even though our EPC includes researchers who are native speakers of several European and Asian languages, we preferentially restrict to English to allow for review and checking of the studies by all team members and also to avoid overburdening nonteam members with translating duties. There was generally consensus among our EPC, our Task Order Officers, and our Technical Expert Panels that including non-English language articles would impose an unnecessary time and resource burden. However, in several instances where the available evidence is of relatively small volume or when we know of important studies in non-English languages, we have gotten data extraction or formal translation done for us.

Table 1. Percentage of studies from Medline in various languages

Language	N	Percent
Total (1996-2011)	2,856,102	100.00%
English	2,620,674	91.76%
Chinese	47,604	1.67%
German	38,357	1.34%
Russian	33,732	1.18%
French	28,550	1.00%
Spanish	26,272	0.92%
Japanese	16,444	0.58%
Italian	10,496	0.37%
Polish	10,018	0.35%
Portuguese	7,440	0.26%
Danish	2,516	0.09%
Dutch	2,202	0.08%
Czech	2,071	0.07%
Hungarian	1,707	0.06%
Norwegian	1,646	0.06%
Turkish	1,482	0.05%
Swedish	1,413	0.05%
Lithuanian	954	0.03%
Korean	950	0.03%
Finnish	914	0.03%
Bulgarian	775	0.03%
Hebrew	711	0.02%
Slovak	436	0.02%
Icelandic	77	<0.01%
Afrikaans	26	<0.01%
Greek	22	<0.01%
Arabic	15	<0.01%
Thai	5	<0.01%
Catalan	3	<0.01%
Macedonian	3	<0.01%
Georgian	2	<0.01%
Malay	2	<0.01%
Indonesian	1	<0.01%
Latvian	1	<0.01%

For a CER we recently conducted, we chose not to apply a language restriction. We ended up including two Spanish language articles. In addition, we needed to review the full text of one French, one Italian, four German, and two Japanese articles. Native German and Japanese speakers were able to screen (and exclude) the latter studies. For the Spanish, French, and Italian studies we tried Google's Web-based translation services, Google Translate® (<http://translate.google.com>). The site can translate large quantities of text that are pasted directly into a text box, or it can be configured to automatically translate foreign language Web pages or PDF files. The program can translate 63 languages (from Afrikaans to Yiddish) into English. Our use of the program was highly successful. The French and Italian language articles were

translated sufficiently clearly for an American with middling French and tourist Italian to be confident about reasons for exclusion. One Spanish article was translated sufficiently clearly for an American with middling Spanish to be confident about the data extraction. The remaining article had one section that seemed to be translated poorly, but a native Spanish speaker confirmed that the original Spanish was just as incomprehensible as the translation.

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. It is the best known such tool among others, including Yahoo!® Babel Fish (babelfish.yahoo.com), Applied Language® (www.appliedlanguage.com/free_translation.shtml), SDL FreeTranslation® (www.freetranslation.com), and Bing® Translator (www.microsofttranslator.com). 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 2550 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 ($\kappa=0.38$) for whether the article should be included in the review.¹⁴

If the translations of the articles are sufficiently accurate for data extraction, the EPCs ought to be able to reliably and easily avoid language restrictions in their reviews. We found only a single article in Medline that considered the use of Google Translate, an editorial focusing on the conceptual problems primary researchers would have translating their manuscripts into English for submission to journals, and advocating for the use of profession translation services.¹⁵

Aims

We conducted a pilot study to formally evaluate the accuracy of one freely available, online, translation tool—Google Translate—for the purposes of data extraction of non-English language articles. We performed simultaneous limited data extraction of a randomly selected convenience sample of recently published non-English language publications and their Google translations. The study was reviewed by the Tufts Health Sciences Campuses Institutional Review Board. It was determined that the activity did not constitute human study research at Tufts University and did not require Institutional Review Board approval.

The research had the following aims:

1. Compare for discrepancies between data extraction done on original-language articles of trials by a native speaker and data extraction done on English-language translations by Google Translate of trials by a researcher who does not know the original article language.

2. Determine the cause of any discrepancies to determine how likely they are to be due to inaccurate translation, and whether there are any clear patterns within, across, or between languages.
3. 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

Study Selection

Based on the frequency of non-English language publications and the languages spoken by native speakers affiliated with the Tufts EPC, we included articles in the following nine languages: Chinese, French, German, Hebrew, Italian, Japanese, Korean, Portuguese, and Spanish. We planned to also include Russian, but we were unable to locate a source of Russian language article PDF or HTML files. The languages were chosen based on a combination of their frequency among articles in PubMed (Table 1) and the availability of past and present Tufts EPC research associates and physician-investigators who are native or fluent speakers of the non-English languages and who have expertise in systematic review and data extraction.

Using QUOSA Information Manager™ (v 8.07.265, QUOSA, Inc.) software, which allowed us to search in PubMed and automatically retrieved available PDF files, we searched with the term “randomized controlled trial,” restricted separately to each of the 10 languages (initially including Russian). This tool can retrieve PDF files from all journals for which the Tufts Health Services Library has a subscription or that are publicly available. We accepted the first 10 publications 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 accepted only studies with these file types since otherwise they could not be translated with Google Translate. Full-text articles were screened by the researcher who was native in that language to determine eligibility. Eligible studies were randomized controlled trials (RCTs) that reported per-treatment group results data (with the exception of Hebrew language studies, see below). 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 RCTs (but were summaries of English-language RCTs). When necessary, we found additional articles from QUOSA to obtain 10 eligible studies per language. When we were unable to find sufficient available trials in a language, the researcher who was native in that language searched country- or journal-specific online databases for relevant studies (e.g., the Korean medical literature database or the Israeli journal *Harefuah*). Upon review of the Hebrew language literature, we found no RCTs in a suitable file format. Therefore, for Hebrew, we included any study that had any comparison between two groups of study participants (whether an intervention or a participant characteristic such as age).

In addition, we chose 10 English-language RCTs to use as a reference standard. These were RCTs that were previously extracted by one of the team members for another systematic review project that included both a continuous and a categorical outcome.

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 did not copy over any English language abstracts that were published with the original articles, but we did copy over English language tables and figures. Each article was translated into a separate Word, PDF, or HTML file that could be accessed without seeing the original article. Translations were performed by the project lead and the research assistant. Where

feasible, we translated articles from languages we could not read. A rough estimate of the time required to extract each study was tracked.

Basic Instructions Compiled for Article Translation

The following are the basic instructions we compiled for internal use to perform article translation. They assume the use of a Microsoft Windows™ operating system. They are not meant to be comprehensive instructions.

1. If you are working from a PDF
 - a. Go to <http://translate.google.com/#>
 - b. Under the large text box, in light blue, click “translate a document”
 - c. Browse to the relevant PDF/HTM.
 - d. Pick the From language.
 - e. Click Translate
2. Save the translation as an HTML file.
3. Google translate seems to maintain the formatting, particularly of tables, much better when it’s working off a Web site (HTM/HTML file) than a PDF document.
 - a. If sections (particularly tables or figures) are not clear, go to the original file and follow the directions in steps 5 & 6 (for those sections or the whole article)
4. If the automatic translation fails
 - a. Copy text (paragraph by paragraph, column by column, or page by page, whichever works cleanly) into a Word document
 - i. Care needs to be taken in some languages (e.g., Hebrew) where the direction of text may be different than English
 - b. Clean up the Word file as necessary (e.g., remove inappropriate line breaks within sentences—particularly for Asian languages, remove hyphens if necessary)
 - c. Copy sections or the whole cleaned up text into the large text box in Google Translate.
 - d. Copy the translated text back into a Word document and save.
 - e. For tables and figures with translatable text (text that can be copied), enter the translations into the appropriate cells in a newly created shell of the table or otherwise indicate which original language text aligns with which translation.
5. If an article consists of blocks of text images (as from scanned documents) for which a machine cannot read lines of text, transformed these images into text by applying an optical character recognition (OCR) process on the file. Then attempt to translate with step 5.
 - a. This approach is likely to work only for languages with Latin alphabets
6. If all translations (or all attempts to copy) from a language fail—particularly those with non-Latin alphabets, you may need to “Install files for complex script and right-to-left languages” or make other modifications to your PC under Regional and Language Options/Language in the Control Panel.

Data Extraction

Data from the original language versions of the articles was extracted by the native speakers. These included two current physician-investigator members of the EPC (French [ID]^{*}, German

^{*} Initials in brackets refer to the study investigators (authors) or acknowledged colleagues.

[KU]), four physician-investigators formerly associated with the EPC (French [GK], Italian and Spanish [JC], Japanese [TT], Portuguese [LZ]), three current EPC research associates (Chinese [MC, WY], Hebrew [NH]), and one former EPC research associate (Korean [JL]). Whenever an article included an English version of the abstract in the original version, extractors of the original language version were instructed to ignore the English version of the abstract.

The English translated versions of the articles were extracted by one of five researchers who did not speak the given language (one physician-investigator [EB] and four research associates [MC, NH, KP, WY]), all currently within the EPC. The extractors of the English language versions were distributed across languages to avoid pairing of original and English language data extractors. Original and English language data extractors were not allowed to review each others' extractions.

With this design, any lack of agreement between the original and English-translated versions can be attributed to either errors in translation or differences between pairs of extractors. To obtain some information on between-extractor variability, the five within-EPC extractors [EB, MC, NH, KP, WY] double-extracted 10 English-language RCTs. Specifically each extracted two English language articles they had previously extracted for a prior systematic reviews and two other English language articles they had never seen before.

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 on those study features that are most important for assessing the study characteristics, methods, and results (see Appendix A for the data extraction form). We limited study quality-related features to objective measures to minimize subjective evaluation of the studies by the data extractors. We extracted the following information: the eligibility criteria, descriptions of the interventions and control, sample size, duration of followup, descriptions and definitions of selected outcomes, the reporting of randomization and allocation concealment techniques, use of blinding, use of intention-to-treat analyses, the reporting of power calculation, and results for selected two outcomes, including baseline value, followup value, mean change, relative effects, confidence intervals, and P values. The selection of outcomes for results data extraction was based on type of data (categorical or continuous), the location of reporting (abstract or full text only), and the completeness of reporting (e.g. mean with standard deviation, per-treatment group data, pre- and post-treatment data). Whenever possible, we selected one categorical outcome and one continuous outcome from each trial, and one outcome that was presented in the abstract (and the full text) and one presented in full text only. We focused on outcomes for which there were direct comparisons between interventions; this approach emphasized effect outcomes. We, thus, mostly excluded adverse events, except where they were reported for all interventions.

The English language extractor was also asked how much additional time was needed to extract translated articles (compared to what it might have taken to extract an equivalent English language article) as “none,” “a little” (up to about a half-hour), or “a lot.”

Data Extraction Comparison

For each study, a single researcher [EMB], with the assistance of a research assistant compared pairs of data extraction forms. The research assistant compared the straightforward pieces of data. The project lead confirmed these and compared extractions of the more clinically or methodologically difficult data (e.g., eligibility criteria, P values). The original plan was to

compare each item in the data extraction form, then for each study, to ask each data extractor to confirm any data from their version of the article for any piece of data for which there was a discrepancy. The pairs of data extractors would then meet to review remaining discrepancies and to come to agreement whether each discrepancy was due to language differences or other reasons. However, four modifications had to be made.

First, the data items from the extraction form were consolidated for the purposes of data comparison (see the annotations in Appendix A). For example, the various types of eligibility criteria asked for were condensed into simply “inclusion criteria” and “exclusion criteria.” Other data items were not analyzed because of lack of relevance or because of wide-ranging disparities in interpretation by the data extractors (e.g., washout period, other blinding methods).

Second, regardless of how many items were extracted, we analyzed (compared) only one intervention, one comparator, the listing of up to five outcomes, the results for one categorical outcome, and one continuous outcome. We chose the first outcomes listed by the original language extractor.

Third, the data reconciliation between data extractors was reduced to simply asking the English-language data extractors (who are all active members of the EPC) to add or confirm data that were missing (compared with the original language extraction) or in the judgment of the project lead required some clarification to assess whether the translation was adequate. In rare instances, the original language extractors (who were mostly off-site) were also asked to fill in missing data; however, in most instances of data missing from the original language extraction, the data item was excluded from the comparison. Exceptions were made, when in the judgment of the project lead the missing data meant “no data” or the English language extraction was sufficiently clear and coherent to be assumed to be accurate. This modification was made both because the volume of data mismatches was so large as to make this step highly time-consuming, and because most of the non-English extractors were off-site (with up to 13 hours time difference), and their availability became limited.

The fourth modification further allowed the researcher doing the data comparison to use his judgment to assess the data extraction forms *in toto* to determine whether there was agreement or not. Examples included making negative inclusion criteria (e.g., not male) to be equivalent to exclusion criteria (female), determining that “no” and “no data” were equivalent, determining whether swapped treatment and comparator was due to arbitrary selection by the extractor or poor translation, and determining whether the P values alternatively extracted as either within or between differences were the same or not. Because of the judgments involved in much of the data comparison, a single researcher (the project lead) made the final comparisons for all studies. This was done to maintain consistency across studies.

Analysis

We calculated the simple percent agreement (items in agreement/total items) as the outcome metric for the analyses. We analyzed percent agreement within sets of studies in each language for each item and for groups of items based on the “tables” on the data extraction form (see Appendix A): eligibility criteria (extraction form table[†] 1; 2 items); intervention and comparator combined (extraction form tables 2a, 2b, 3, 3a, and 3b; 12 items); design (extraction form table 5; 4 items), quality issues (extraction form table 6; 9 items); outcomes (extraction form table 7; 7 items); categorical results (extraction form table 8; 9 items); and continuous results (extraction

[†] These tables refer to the “tables” in the data extraction form, not the Tables in this report.

form table 9; 27 items). Histograms of the percent agreement for all items together and for each category group within each language (including English) were graphed so that comparisons could be made across languages. The English language study comparisons acted as a reference standard to compare the degree of agreement we achieved by extracting data from English language articles with the degree of agreement for each language. We did not use kappa statistics because the large majority of items were not dichotomous. In general, we were comparing descriptions (e.g., “inclusion criteria”).

We first performed Mann-Whitney tests to compare the distribution of agreements across all extraction items for each foreign language (separately) and English language extraction. We repeated the same test for each category of items between each foreign language and English language. Based on the observed distribution of our reference standard (i.e., English language), we defined “good agreement” as greater or equal to 80 percent agreement. We performed the Fisher’s exact test to assess the differences in the percentage of items that reached “good agreement” between each foreign language and English language, across all categories, for each category of items, and for each language set of studies (the percentage of studies that had >80 percent agreement within each study).

Analyses were conducted with Stata SE 11 software (Stata Corp., College Station, Texas). All P values were 2-tailed, and a P value less than 0.05 was considered to indicate a statistically significant difference. We did not adjust for multiple testing. The researcher performing the comparisons also collected examples of obvious causes of disagreements between original language and English extractions.

Results

Study Selection

As described in the Methods section, we originally planned to include 10 RCTs from each of 10 languages (in addition to English). We had to drop Russian since we could not locate a source of Russian language studies in PDF or HTML file format. For Hebrew, we had only one source of studies in PDF or HTML file format (the journal *Harefuah*); however none of the files that could be translated included RCTs. Furthermore, we found only eight Hebrew language studies that compared two groups of study participants. Thus we analyzed 88 non-English articles (plus 10 English RCTs); see Appendix B for the list of articles utilized.

Article Translation

Using Google Translate we were unable to translate 21 articles that met eligibility criteria. These included one each in French, German, and Japanese (1/11, 9 percent), two in Chinese (2/12, 17 percent), three in Korean (3/13, 23 percent), four in Italian (4/14, 29 percent), nine in Hebrew (9/17, 53 percent), but none in Portuguese and Spanish. The failures occurred because Google Translate could not read the PDF or HTML file, and we could not copy out of the PDF file, optical character recognition failed, or Google Translate (or our computers) could not recognize the letters or characters (this occurred most commonly with the Asian languages).

The length of time required to translate articles ranged from seconds (51 of 88 articles, 58 percent) to about 1 hour. Using a rough average of 15 seconds for the articles for which it was coded to take “seconds” to translate, the average time to translate was about 10 minutes. The time-for-translation distributions varied by language (Table 2).

Table 2. Translation time, by language

Articles*:	1	2	3	4	5	6	7	8	9	10
European										
French	secs	30 min	45 min	60 min						
German	secs	20 min	20 min							
Italian	secs	1 min								
Portuguese	secs	45 min								
Spanish	secs	15 min								
Asian										
Chinese	secs	secs	5 min	10 min	10 min	10 min	10 min	10 min	20 min	45 min
Japanese	secs	20 min	45 min	60 min						
Korean	10 min	15 min	20 min	45 min	45 min	45 min				
Other										
Hebrew	--	--	30 min	30 min	30 min	30 min	40 min	45 min	45 min	50 min

min = minute(s); secs = “seconds” (the translation was quick, approximately <1 minute).

* For each language, the duration of time for translation of each article is listed, sorted from shortest to longest time. Those articles coded as taking “seconds” to translate are shaded blue. Those articles that took from 1 to 20 minutes to translate are shaded green. Those that took longer (30-60 minutes) are shaded orange.

In general, the European and Japanese language articles could be translated automatically from their PDF or HTML files without manipulation by the research assistant prior to translation. However, the ease of translation was largely related to the file/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 mainly consisted 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. We discovered (and were informed by the Chinese speakers among us) that we also needed to remove false line breaks (artificial 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 Hebrew and Asian languages, Google Translate could directly translate the PDF or HTML file, but the resulting file was unreadable because of 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 these 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. Appendix C includes examples of poorly translated articles, including some with modest amounts of overlap. Other issues that we encountered included that Google Translate failed to translate an Italian article on one day but succeeded on a later day; one Korean article could not be read on one computer, but could on another computer; one Spanish PDF could not be read originally but could after it was saved as a tiff file from which another PDF was created; and one German article required removing multiple instances of “-” (an optional hyphen) before translation could succeed.

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. (We did not directly compare extraction times since this would have compared extraction speeds of different extractors more than added extraction time due to translation.) For only six (7 percent) studies did the English extractor think that no extra time was needed because of the translation (one Italian, two Korean, one Portuguese, two Spanish); however, one data extractor may have been much more forgiving than others, since five of these six were extracted by one researcher. For 40 articles (45 percent) “a little” extra time was required and for 42 (48 percent) “a lot” of extra time was required. The languages requiring the most extra time to extract (at least half the articles required “a lot” of extra time), in order of extra time required were Chinese (80 percent “a lot”), German (70 percent), Japanese (70 percent), Korean (50 percent), Hebrew (50 percent). However, for certain Hebrew papers (and possibly papers in other languages), little extra time was needed since the translation was so poor that the extractor quickly determined that data could not be extracted. The four Romance languages (Italian, French, Portuguese, and Spanish) required the least extra time to extract.

Google Translate frequently translated non-European language text into gibberish (e.g., “Ahpthofizivlogy,” “Gyeongjeongmaekgan”) or nonsense text (e.g., “cantharidin poisoning attack erosion are sore,” “the poet tested discretely”).

There was general agreement that data described in the text (e.g., eligibility criteria, study methodology) were much more difficult to extract than data reported in tables and figures (i.e., most results). Descriptions of interventions other than drugs and placebo were much more difficult to extract than drugs.

Many of the Hebrew articles posed a particular challenge, because Google Translate often flipped the direction of numbers. (Though Hebrew is written from right to left, numbers are written from left to right as in English; however, Google Translate usually flipped the direction of all items.) Similarly, the translation of the Asian languages often jumbled sentences with numbers in them such that it could not be determined which numbers matched which items (e.g.,

“The laboratory values of WBC $\geq 3,000 / \text{mm}^3$, $i \geq$ platelets Number $10 \times 10^4 / \text{mm}^3$, $6 \geq$ total protein. O g/dL (A/G ≥ 1.0), AST, ALT $100 \leq, \leq$ serum creatinine 1.5 mg/d.”

Comparison of Translated With Original Articles

Figures 1–8 show histograms for each language of the distribution of percent agreement of all extracted items (Figure 1) and of items within different sections (Figures 2–8). For each histogram, the median and interquartile range (IQR) of percent agreement is displayed, along with the percent of items that had ≥ 80 percent agreement (in the upper right corner of each histogram). The European languages are grouped together in the upper section and the Asian languages in the lower section. The histograms for English articles are displayed separately at the bottom of each figure to use as a reference standard.

We evaluated up to 70 extracted items per article (see below for items). The actual number of compared items per article ranged from 10 to 65, with a mean of 39 and a median (IQR) of 41 (34, 45) items. We arbitrarily define “high agreement” to mean there was at least 80 percent agreement within an item or article.

Evaluating all items analyzed together (Figure 1), 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. This compared with high agreement in English language agreement for 80 percent of the items. French, Italian, Japanese, and Korean articles had high agreement for about 40 to 45 percent of items. Spanish articles for 30 percent of items, Hebrew articles for 24 percent of items, and Chinese for 8 percent of items. Table 3 shows that the absolute agreement and the percent of items with high agreement were statistically significantly worse for all languages than for English articles. Furthermore, 8 of 10 English language articles had high agreement for all items; compared with similar levels of agreement among Portuguese and German articles, lesser agreement among other languages, and notably, no Chinese articles with high agreement.

Table 3. Agreement across all items, by language

Language	Agreement, Median (IQR)	P*	Items with $\geq 80\%$ Agreement	P*	Articles with $\geq 80\%$ Agreement	P*
Portuguese	0.85 (0.73, 1.00)	0.015	40/64 (63%)	0.032	7/10 (70%)	1.00
German	0.82 (0.63, 1.00)	0.001	37/60 (62%)	0.029	6/10 (60%)	0.63
French	0.75 (0.67, 1.00)	<0.001	31/70 (44%)	<0.001	4/10 (40%)	0.17
Korean	0.75 (0.63, 1.00)	<0.001	27/59 (46%)	<0.001	4/10 (40%)	0.17
Japanese	0.75 (0.50, 0.83)	<0.001	29/64 (45%)	<0.001	3/10 (30%)	0.070
Italian	0.70 (0.60, 0.90)	<0.001	26/63 (41%)	<0.001	4/10 (40%)	0.17
Spanish	0.67 (0.50, 0.80)	<0.001	18/61 (30%)	<0.001	3/10 (30%)	0.070
Hebrew	0.60 (0.00, 0.75)	<0.001	9/37 (24%)	<0.001	3/8 (37.5%)	0.15
Chinese	0.50 (0.25, 0.57)	<0.001	5/65 (8%)	<0.001	0/10 (10%)	0.001
English	1.00 (0.80, 1.00)	ref	53/66 (80%)	ref	8/10 (80%)	ref

ref = reference language for statistical test (P value)

*P value versus English

Table 4 displays the list of sections and their respective items that we next analyzed. Tables 5 and 6 display the absolute agreement (Table 5) and the percent of items with high agreement (Table 6) by each section of the data extraction form. Similar patterns hold across languages. Despite the subjective assessment by the data extractors that certain sections were more difficult to extract than others, there were no consistent patterns evident across sections regarding the degree of agreement.

Table 4. Analyzed data extraction items, by section

Section	n	Items		
Eligibility criteria	2	Inclusion criteria	Exclusion criteria	
Description of treatment / control	12	Treatment name	Control name	
		Treatment description	Control description	
		Treatment dose	Control dose	
		Treatment frequency	Control frequency	
		Treatment route	Control route	
Study characteristics	4	Treatment duration	Control duration	
		Followup duration	Number enrolled	
Study methodology	9	Number of centers	Number analyzed	
		Blinded patient (Y/N)*	Randomization technique reported (Y/N)*	
		Blinded caregiver (Y/N)*	Allocation concealment method reported (Y/N)*	
		Blinded outcome assessment (Y/N)*	Claimed intention-to-treat (Y/N)*	
		Reported “double blind” (Y/N)*	Power calculation reported (Y/N)**	
Outcomes description	7	Reported “single blind” (Y/N)*		
		Outcome name (1st outcome)	Outcome name (3rd outcome)	
		Outcome description (1st outcome)	Outcome name (4th outcome)	
		Outcome name (2nd outcome)	Outcome name (5th outcome)	
Dichotomous results	9	Outcome description (2nd outcome)		
		No. events (treatment arm)	Result / value	
		No. total (treatment arm)	95% confidence interval	
		No. events (control arm)	P value (difference between arms)	
Continuous results	27	No. total (control arm)	Factors adjusted for	
		Metric reported (e.g., odds ratio)		
		Unit of measurement	Measure of variability† (baseline & final)	Value change (treatment)
		No. baseline (treatment)	No. baseline (control)	Variability change (treatment)
		Value baseline (treatment)	Value baseline (control)	P value change (treatment)
		Variability baseline (treatment)	Variability baseline (control)	Value change (control)
		No. final (treatment)	No. final (control)	Variability change (control)
		Value final (treatment)	Value final (control)	P value change (control)
		Variability final (treatment)	Variability final (control)	Factors adjusted for
Value net difference	Measure of variability† (within-arm change)	Between-arm difference type‡		
Variability net difference	Measure of variability† (between-arm difference)	P value (between-arm difference)		

* Y/N = yes/no question.

† Standard deviation, standard error, or 95% confidence interval

‡ Difference between final values or net difference

Table 5. Percent of items for which there was agreement, by language and data extraction form section

Section :	Eligibility (n=2)	Tx/Cx (n=12)	Study Char (n=4)	Metho ds (n=9)	Outcomes (n=7)	Dich Res (N=9)	Cont Res (n=27)
Language	Media n (IQR)*						
	P †	P †	P †	P †	P †	P †	P †
Portuguese	0.85 (0.70, 1.00)	0.78 (0.71, 0.92)	0.90 (0.79, 1.00)	1.00 (0.90, 1.00)	0.88 (0.71, 1.00)	0.90 (0.50, 1.00)	0.80 (0.67, 1.00)
German	0.80 (0.70, 0.90)	1.00 (0.82, 1.00)	1.00 (0.94, 1.00)	0.90 (0.80, 0.90)	0.63 (0.57, 0.80)	0.57 (0.57, 0.57)	0.80 (0.60, 1.00)
French	0.94 (0.89, 1.00)	0.80 (0.69, 0.92)	0.79 (0.74, 0.85)	0.80 (0.80, 0.90)	0.67 (0.67, 1.00)	0.75 (0.67, 1.00)	0.67 (0.50, 1.00)
Korean	0.65 (0.50, 0.80)	0.83 (0.75, 1.00)	0.70 (0.65, 0.75)	0.80 (0.70, 0.90)	0.71 (0.33, 1.00)	1.00 (1.00, 1.00)	0.67 (0.50, 0.75)
Japanese	0.42 (0.33, 0.50)	0.71 (0.67, 0.79)	0.45 (0.37, 0.58)	0.90 (0.90, 0.90)	0.63 (0.33, 1.00)	0.80 (0.50, 1.00)	0.75 (0.50, 0.80)
Italian	0.76 (0.67, 0.86)	0.69 (0.60, 0.89)	0.79 (0.74, 0.90)	0.70 (0.60, 0.90)	0.89 (0.60, 1.00)	0.00 (0.00, 1.00)	0.67 (0.50, 1.00)
Spanish	0.78 (0.75, 0.80)	0.83 (0.75, 1.00)	0.84 (0.71, 0.94)	0.70 (0.50, 0.70)	0.80 (0.63, 1.00)	0.67 (0.50, 0.67)	0.50 (0.00, 0.60)
Hebrew	0.75 (0.75, 0.75)	0.68 (0.55, 0.78)	0.71 (0.65, 0.75)	‡ ‡	0.38 (0.00, 1.00)	0.63 (0.60, 0.90)	0.00 (0.00, 0.50)
Chinese	0.42 (0.33, 0.50)	0.42 (0.27, 0.68)	0.56 (0.50, 0.67)	0.56 (0.44, 0.67)	0.75 (0.44, 0.89)	0.57 (0.57, 0.57)	0.25 (0.00, 0.50)
English	0.85 (0.80, 0.90)	1.00 (1.00, 1.00)	0.90 (0.80, 0.95)	0.90 (0.80, 1.00)	1.00 (1.00, 1.00)	0.87 (0.70, 1.00)	1.00 (0.72, 1.00)

“n” refers to the number of items per section.

ref = reference language for statistical tests (P value); Tx/Cx = treatment and control; Study Char = study characteristics; Dich Res = dichotomous results; Cont Res = continuous results.

Red: $P \leq 0.05$; yellow: $0.05 < P < 0.10$; blue: $P \geq 0.10$.

* Percent agreement among items in section

† P value versus English

‡ These items, related to randomized controlled trials, were not extracted for Hebrew observational studies.

Table 6. Items with at least 80 percent agreement, by language and data extraction form section

Section:	Eligibility (n=2)	Tx/Cx (n=12)	Study Char (n=4)	Methods (n=9)	Outcomes (n=7)	Dich Res (N=9)	Cont Res (n=27)
Language	≥80%* P † #/N (%)	Median (IQR)* P †					
Portuguese	1/2 (50%) 1.00	6/12 (50%) 0.069	3/4 (25%) 1.00	8/9 (11%) 1.00	5/7 (29%) 0.46	5/8 (38%) 1.00	12/22 (45%) 0.55
German	1/2 (50%) 1.00	11/12 (8%) 1.00	4/4 (0%) 1.00	7/9 (22%) 1.00	2/7 (71%) 0.021	0/5 (100%) 0.021	12/21 (43%) 0.55
French	2/2 (0%) .	7/12 (42%) 0.16	2/4 (50%) 1.00	7/9 (22%) 1.00	3/7 (57%) 0.070	3/9 (67%) 0.15	7/27 (74%) 0.005
Korean	1/2 (50%) 1.00	7/11 (36%) 0.16	1/4 (75%) 0.49	5/9 (44%) 0.29	3/7 (57%) 0.070	5/5 (0%) 0.49	5/21 (76%) 0.007
Japanese	0/2 (100%) 0.33	3/12 (75%) 0.003	0/4 (100%) 0.14	9/9 (0%) 1.00	2/7 (71%) 0.021	5/9 (44%) 0.62	10/21 (52%) 0.24
Italian	1/2 (50%) 1.00	4/10 (60%) 0.020	2/4 (50%) 1.00	3/9 (67%) 0.050	4/7 (43%) 0.19	4/9 (56%) 0.34	8/22 (64%) 0.075
Spanish	1/2 (50%) 1.00	8/12 (33%) 0.32	3/4 (25%) 1.00	2/9 (78%) 0.015	4/7 (43%) 0.19	0/5 (100%) 0.021	0/22 (100%) <0.001
Hebrew	0/2 (100%) 0.33	1/4 (75%) 0.027	0/4 (100%) 0.14	‡ ‡	2/6 (67%) 0.021	3/8 (63%) 0.32	3/13 (77%) 0.017
Chinese	0/2 (100%) 0.33	1/12 (92%) <0.001	0/4 (100%) 0.14	0/9 (100%) <0.001	3/7 (57%) 0.070	0/5 (100%) 0.021	1/26 (96%) <0.001
English	2/2 (0%) ref	11/12 (8%) ref	3/4 (25%) ref	8/9 (11%) ref	7/7 (0%) ref	6/8 (25%) ref	16/24 (33%) ref

“n” refers to the number of items per section.

ref = reference language for statistical tests (P value); Tx/Cx = treatment and control; Study Char = study characteristics; Dich Res = dichotomous results; Cont Res = continuous results.

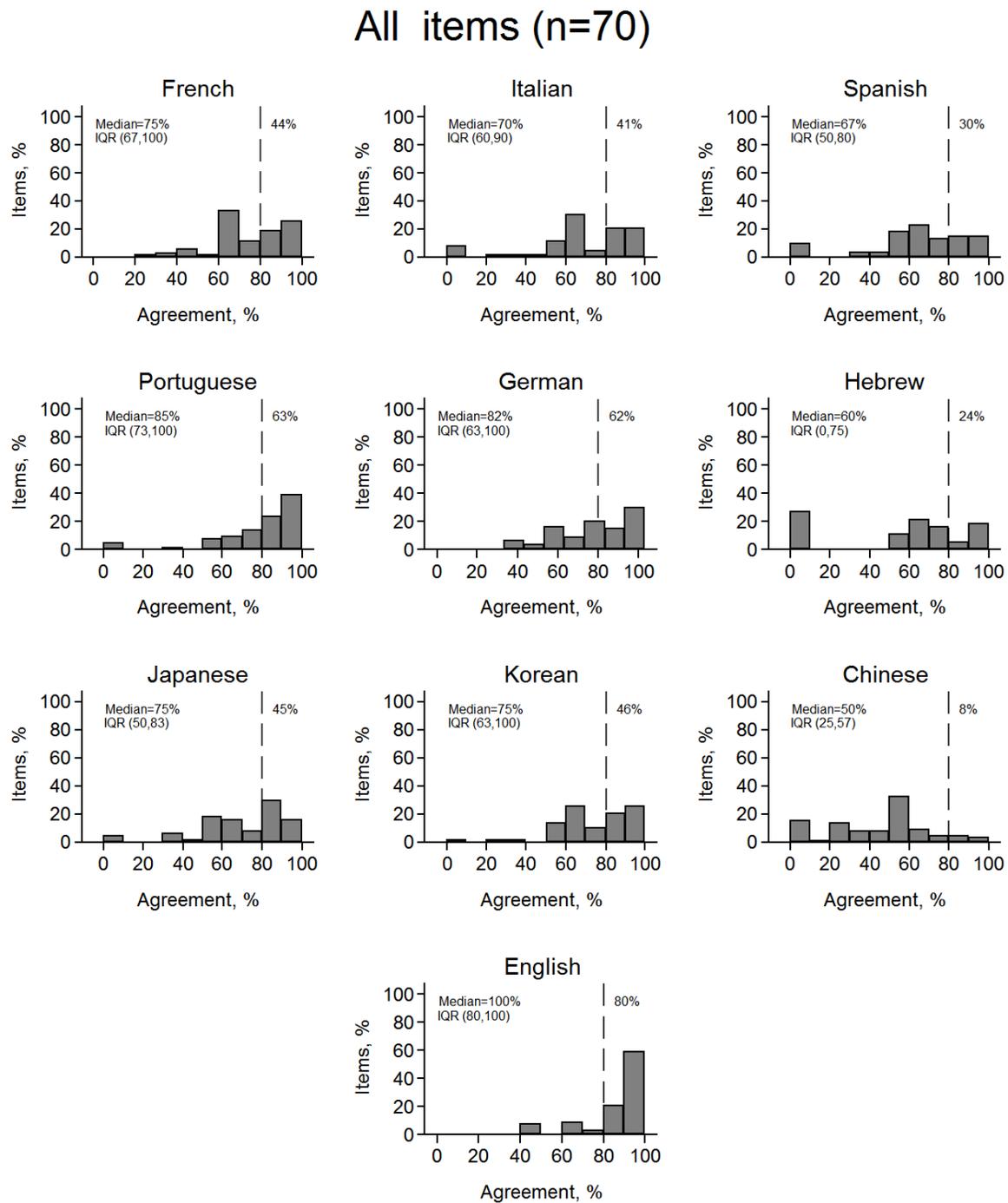
Red: P<0.05; yellow: 0.05<P<0.10; blue: P≥0.10.

* Number and percent of items that had ≥80% agreement in section

† P value versus English

‡ These items, related to randomized controlled trials, were not extracted for Hebrew observational studies.

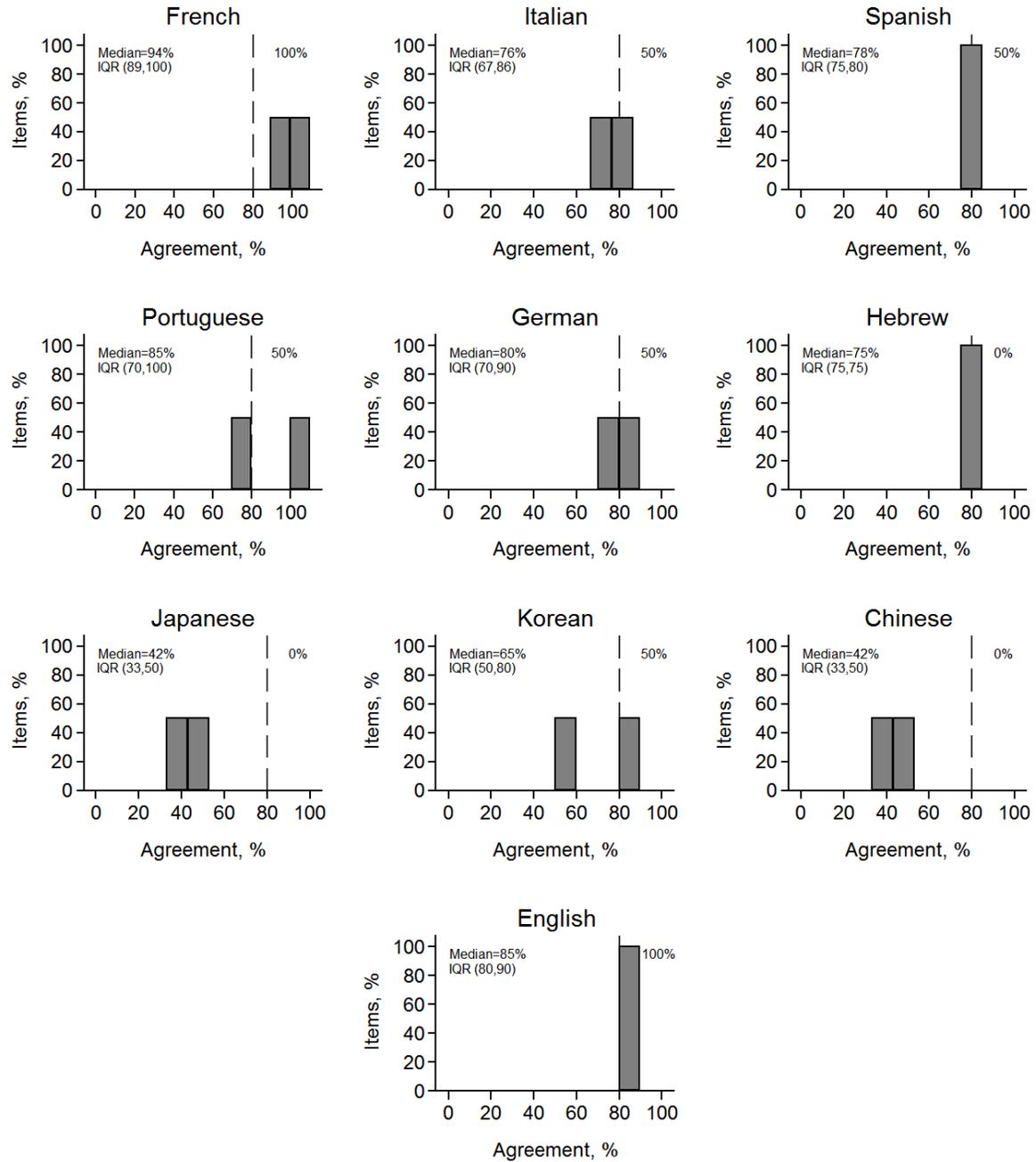
Figure 1. Histograms of the percent agreement for all items, by language



There were a maximum of 70 items per study. All histograms are drawn on the same scale, with 10 percentage point wide bars. For each histogram (language), the median and interquartile range (IQR) percent agreement across the 10 trials (8 observational studies in Hebrew) are displayed in the upper left corner. The percentage figure displayed in the upper right corner of each histogram represents the percentage of items for which there were >80 percent agreement (indicated by the vertical dashed line).

Figure 2. Histograms of the percent agreement for eligibility criteria items, by language

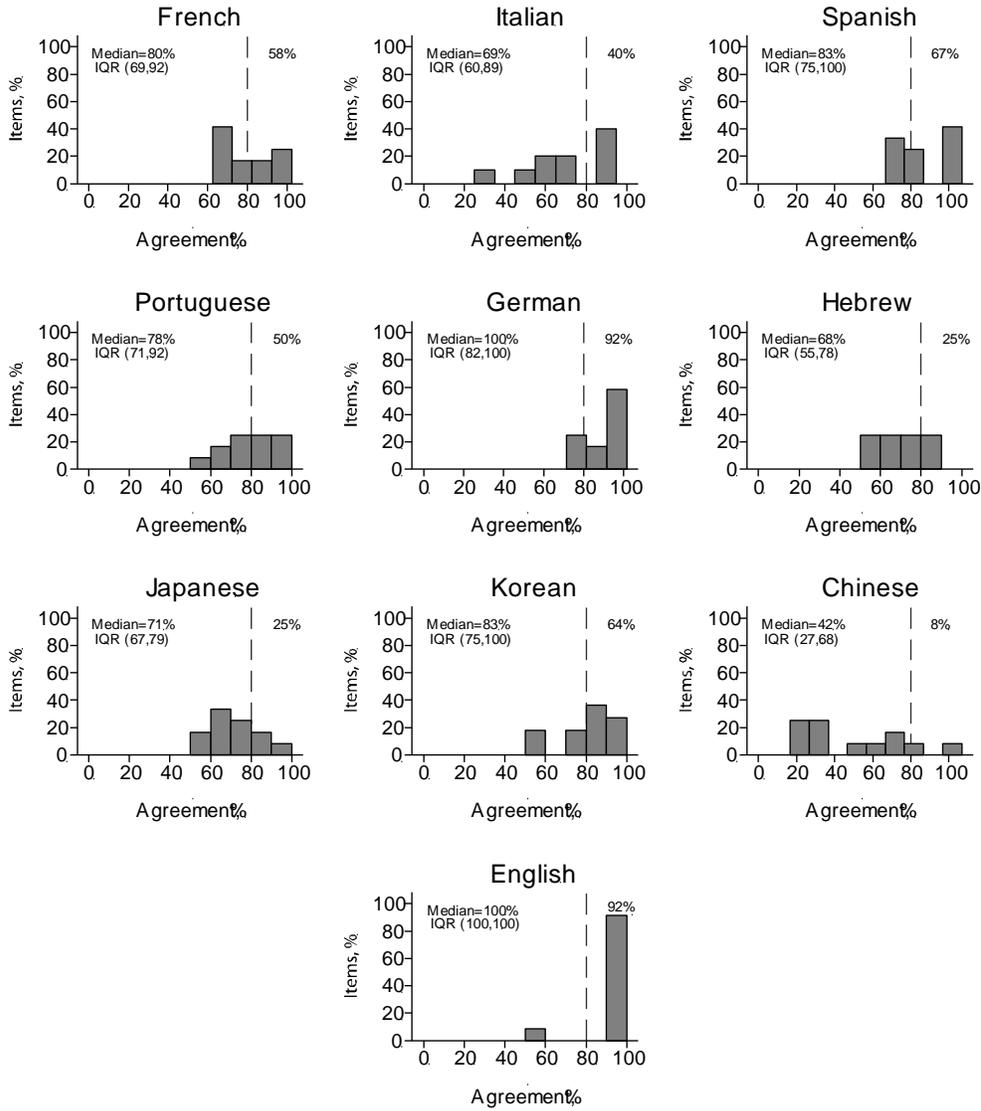
Eligibility criteria items (n=2)



There were a maximum of two such items per study. See legend to Figure 1 and Table 4.

Figure 3. Histograms of the percent agreement for descriptions of treatment and control items, by language

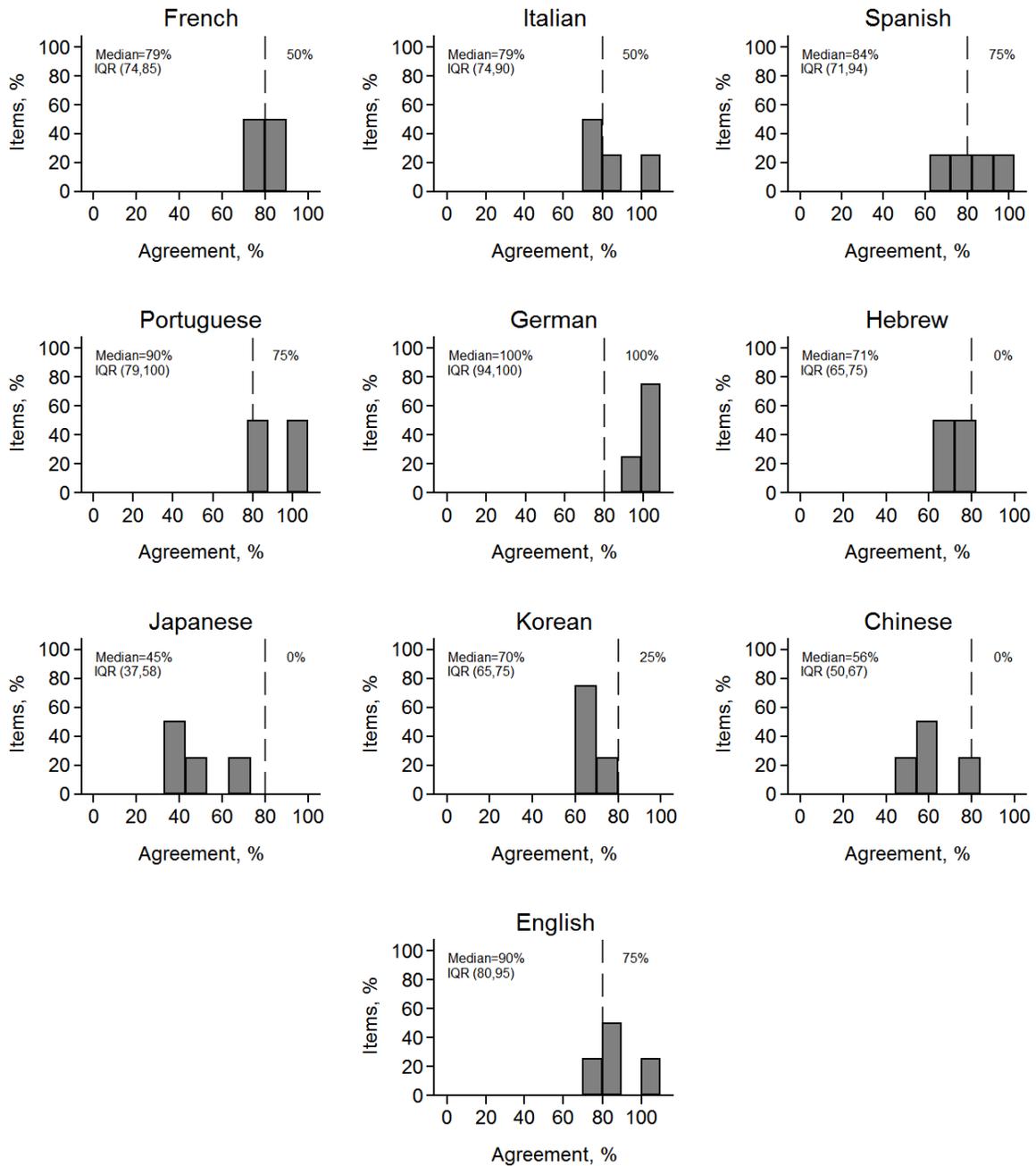
Description of treatment/control items (n=12)



There were a maximum of 12 such items per study. See legend to Figure 1 and Table 4.

Figure 4. Histograms of the percent agreement for study characteristics items, by language

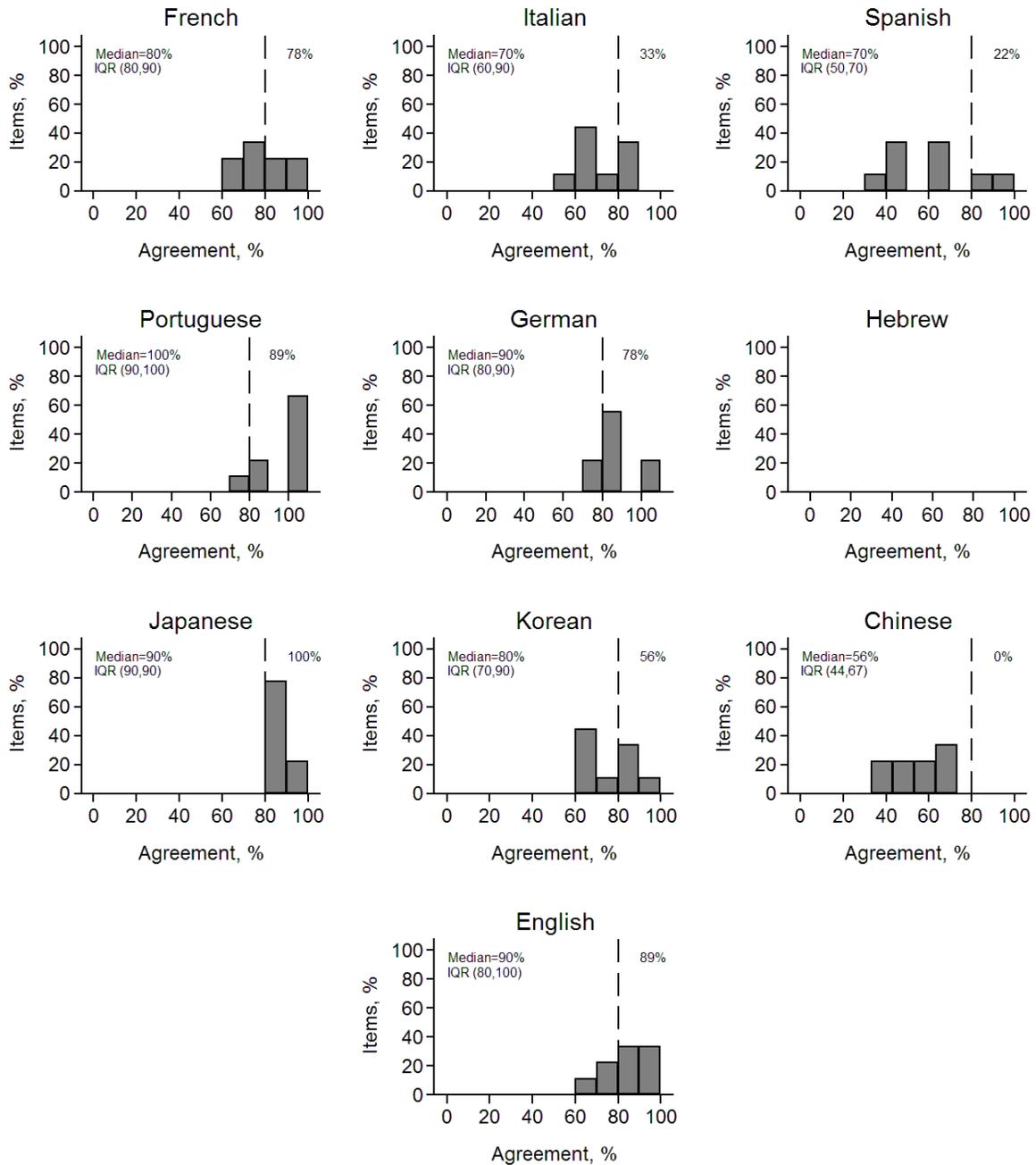
Study characteristics items (n=4)



There were a maximum of 4 such items per study. See legend to Figure 1 and Table 4.

Figure 5. Histograms of the percent agreement for study methodology items, by language

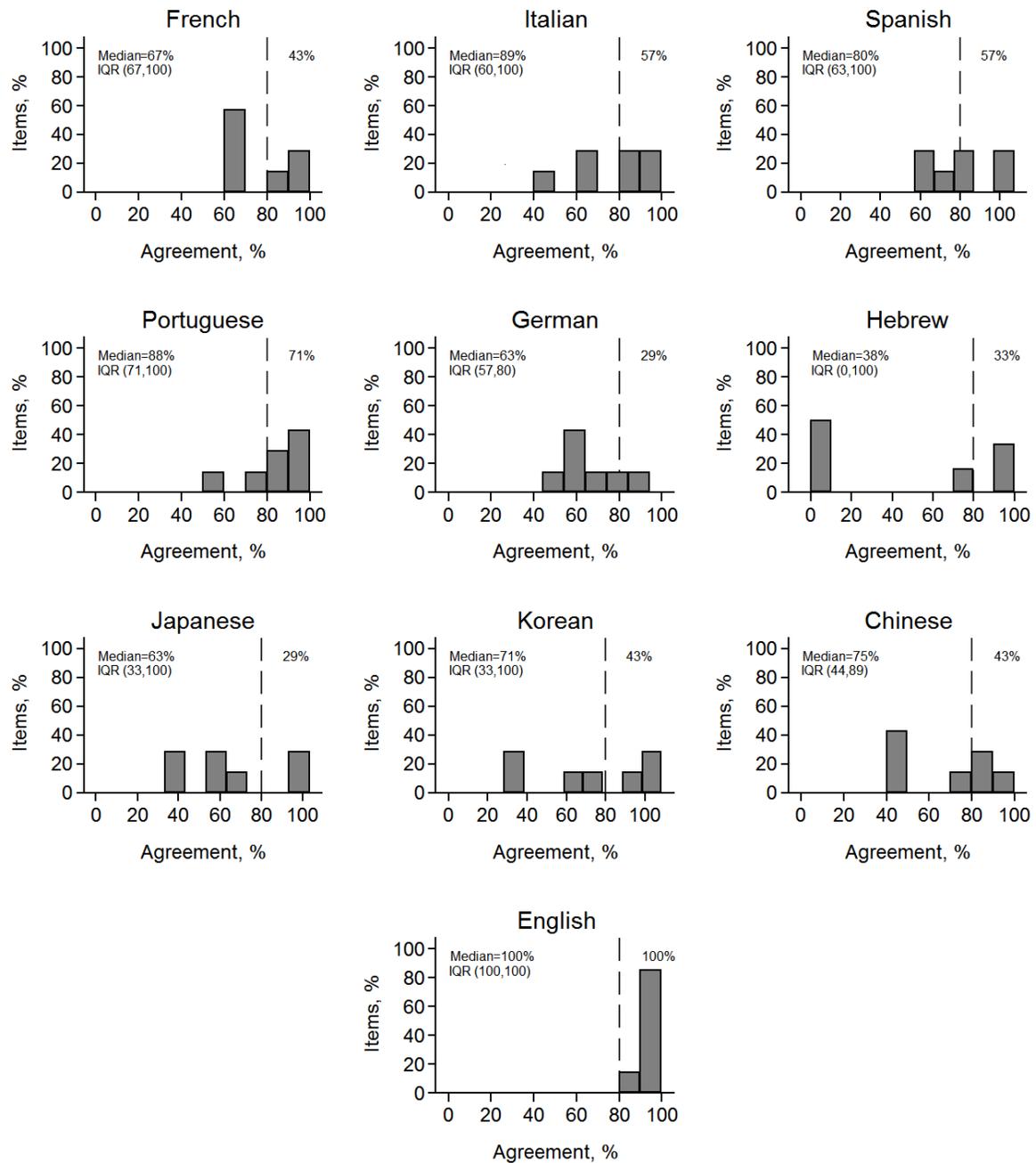
Study methodology items (n=9)



There were a maximum of 9 such items per study. See legend to Figure 1 and Table 4.

Figure 6. Histograms of the percent agreement for descriptions of outcomes items, by language

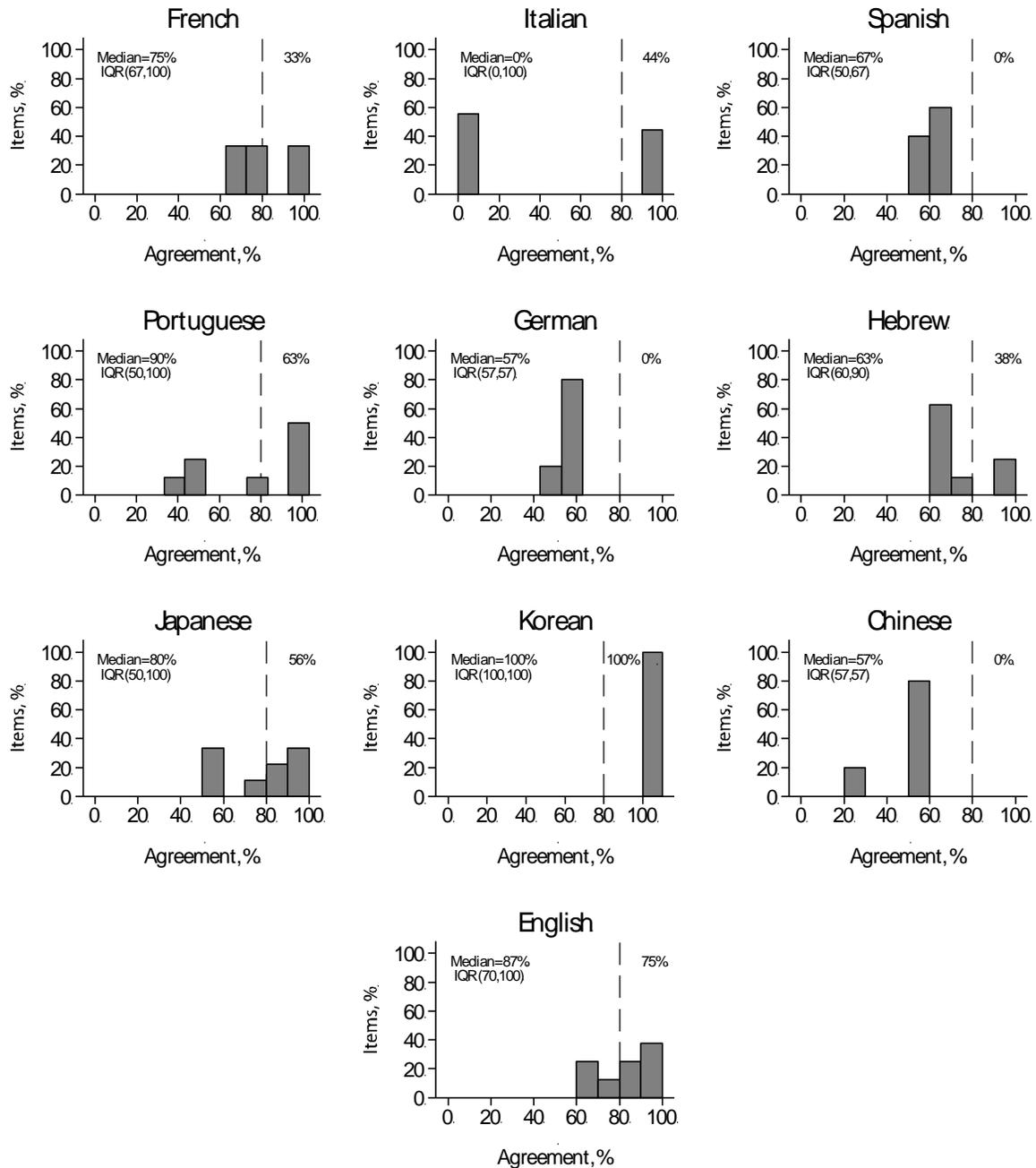
Description of outcomes items (n=7)



There were a maximum of seven such items per study. See legend to Figure 1 and Table 4.

Figure 7. Histograms of the percent agreement for categorical results items, by language

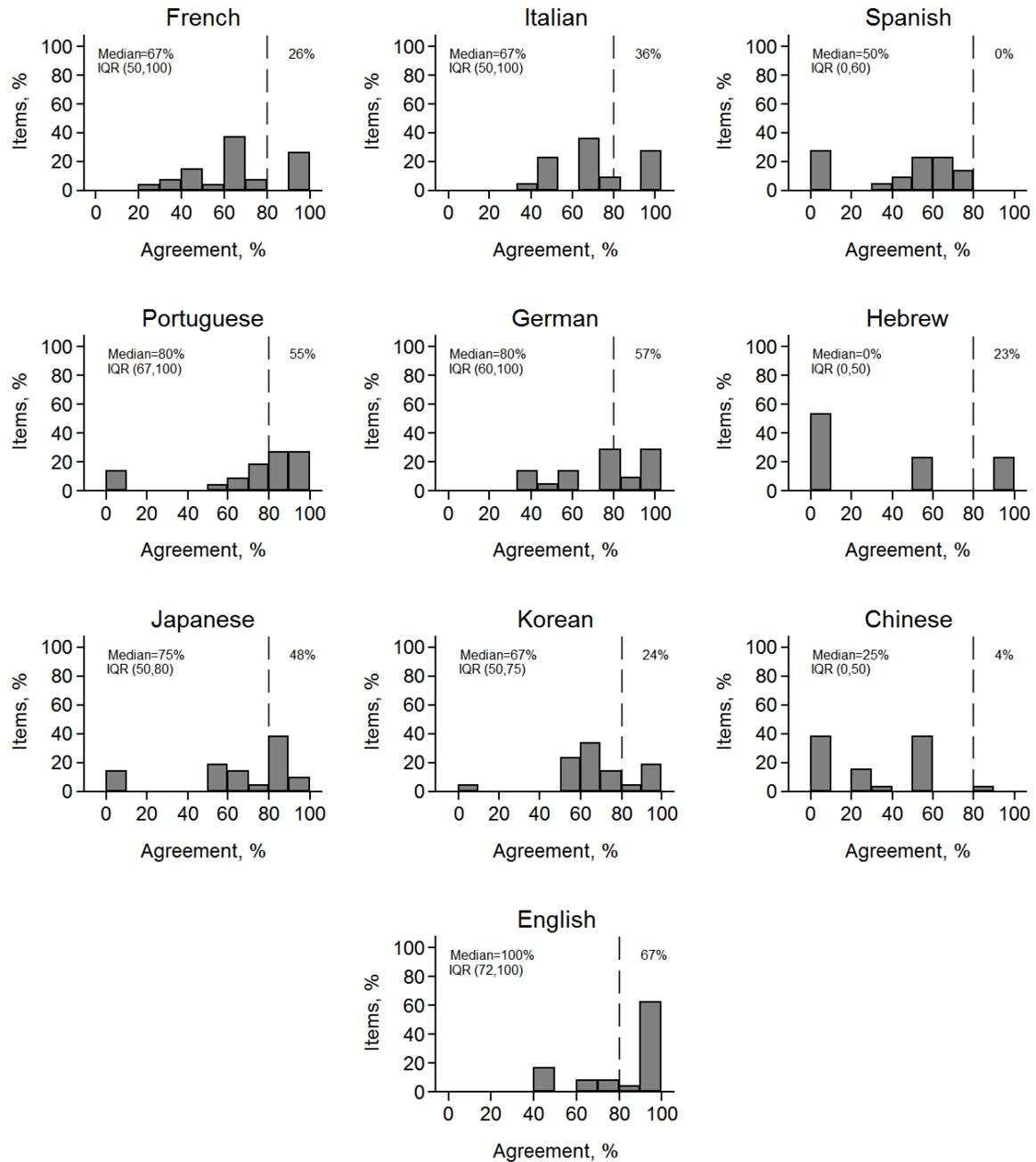
Dichotomous results items (n=9)



There were a maximum of nine such items per study. See legend to Figure 1 and Table 4.

Figure 8. Histograms of the percent agreement for continuous results items, by language

Continuous results items (n=27)



There were a maximum of 27 such items per study. See legend to Figure 1 and Table 4.

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, Romance languages had much higher levels of agreement than Asian languages and Hebrew. This difference across languages was similar to the findings of machine translation experts for general translation.¹³

This study had several important limitations. This was a pilot study with a small number of articles extracted in each language, with only a single extractor available for each language. (We did have two extractors for French, but they did not review each other's extractions.) We were unable to fully ascertain the accuracy of the data extraction in the original languages, which may have resulted in spuriously high rates of disagreement. We were not able to directly attribute disagreement between extractors to errors in translation, as disagreement could also be due to different extractors interpreting articles in different ways or errors in extraction. This effect may have been extractor-dependent, which would have manifested as being language-dependent.

We could not confirm the accuracy of extractions, particularly those done from the original language articles. Data extractors of the translated articles were asked to fill in missing or unclear data, but we were unable to coordinate full determination of why there were disagreements. Therefore, we added a double-extraction of English language articles to use as a reference standard to gauge the degree of disagreement in the translated articles. In addition, 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.

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. Furthermore, the assessment of whether extractors agreed with each other was inherently subjective for many items.

In addition, our grouping of extraction items into sections was arbitrary and we did not adjust for the relative size or complexity of items within each study. Thus, it might not be appropriate to directly compare across sections. Percent agreement for eligibility criteria (with two items) is conceptually different than percent agreement for continuous outcomes (with 27 items). Thus, the apparent oddity that "results" sections (dichotomous and continuous) had low levels of agreement (e.g., a median of 25 percent for Chinese) may be largely explained by the number and breadth of specific items within these sections.

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. Tricks for more rapid and more accurate translation can also easily be gathered and made available to all the EPCs. 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. The exception to this was when articles were so badly translated that it was clear that little usable data could be extracted. This occurred most commonly with Chinese language articles and also with Hebrew articles.

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 for most of the tested languages it may be worth attempting to translate (with Google Translate) and extract non-English language articles that are available as machine-readable PDF (or HTML) files. Based on the fluency and legibility of the translation, the reviewer should be able to make an assessment regarding how much confidence to have in the accuracy of the translated version. It may 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 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. Although we ranked the languages by agreement, based on statistical analyses, we do not claim that the exact ranking truly represents the actual level of accuracy one could expect from future data extractions of translated articles. However, it may be fair to say that one can expect fair to good translation (for the purposes of data extraction) from European languages, fair translation from Japanese and Korean; but often poor translation from Chinese and Hebrew.

Before the systematic review community can be confident in the value of using Google Translate to allow inclusion of non-English language articles, more research is needed to explore its value and its limitations. A future evaluation could focus on specific languages and possibly on a narrower or simpler list of data extraction elements. A followup study should also perform double (or more) data extraction for both the original language and the translated articles, or otherwise control for innate differences among extractors; this would allow a better determination that lack of agreement in extraction are due to translation errors, rather than differences among extractors. This investigation focused on Google Translate as a translation tool, but there are other online, free translation programs that should also be tested, and possibly compared. A formal collaborative study by the EPCs could harness the language skills across the different centers, would enable multiple duplication of data extractions, and would improve the generalizability of experiences of extracting translated articles beyond those of a single EPC.

We conclude that more research is necessary to better understand the utility of this new translation tool to reduce the risk of language bias in systematic review. However, in the meantime, it may be worthwhile for EPCs 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. Investigating Google Translate (or other Web-based translation tools) as a collaborative research project across all EPCs would take advantage of the quick accrual rate of a multicenter study as well as the benefits of a prospective study design. It

would allow for coordination of the various centers' experiences with using translated articles in reports and will reduce the anecdotal nature of a single EPC's experiment.

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. National Research Council. *Finding What Works in Health Care: Standards for Systematic Reviews*. Washington, DC: The National Academies Press; 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, Hoenstein 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.
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)*http://findarticles.com/p/articles/mi_7099/is_4_9/ai_n56337599/?tag=content. 2009. Accessed January 9, 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 January 6, 2012.
14. Freitas de Souza R, Sequeira P, Nasser M, et al. 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>. 2009. Accessed January 30, 2012.
15. [Sheppard F](#). Medical writing in English: The problem with Google Translate. *Presse Med*. 2011 Jun;40(6):565-6. Epub 2011 Apr 22. PMID: 21514783.

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
EPC	Evidence-based Practice Center
IOM	Institute of Medicine
IQR	interquartile range
OCR	optical character recognition
RCT	randomized controlled trial

Appendix A. Annotated Data Extraction Forms

The annotations (comments) are included to explain how the extracted data were used in the analyses of agreement across extractors.

Keep extracted data concise. Recognize that this is a generic form, not topic-specific.³

Author, Year		PMID	
Extractor		RefID	
Language			

1. Eligibility criteria

Inclusion criteria: demographics	
Inclusion criteria: disease/conditions (generic, e.g. HTN)	
Inclusion criteria: disease/conditions (specifics, e.g. SPB >140)	
Inclusion criteria: other ⁴	
Exclusion criteria: comorbidity	
Exclusion criteria: other ⁵	

2. Interventions⁶

2(a) If interventions are drugs/supplements (other interventions that fit):

	Intervention drug name	Dose	Frequency	Route	Duration of intervention
1					
2					
3					

2(b) If interventions are not drugs^{*}

	Intervention name	Concise Description	Frequency	Duration of intervention
1				
2				
3				

* If a cointervention (eg, education) is used in all patients, enter info in 4, not here.

If an intervention has multiple components (that are all different than in other study arms) enter each on a separate row and renumber the 1st column as needed.

3. Comparator

Type of comparator: (other drug, placebo, usual care, no treatment, etc.)	
--	--

3a. If comparator is a drug/supplements (other interventions that fit):

	Comparator drug name	Dose	Frequency	Route	Duration of intervention
1					
2					
3					

3b. If comparator is not a drug:

	Comparator name	Concise Description	Frequency	Duration of intervention ⁸
1				
2				
3				

³ The comments below indicate how extracted data were handled during reconciliation (comparison of two data extraction forms).

⁴ Grouped together as "Inclusion Criteria".

⁵ Grouped together as "Exclusion Criteria".

⁶ Simplified to 6 data: Intervention name; Dose; Concise description; Frequency; Route; Duration of intervention.

⁷ Excluded from analysis.

⁸ Simplified to 6 data: Comparator name; Dose; Concise description; Frequency; Route; Duration of intervention.

4. Co-interventions⁷

	Co-intervention name	Description (include dose, frequency, and other details)
1		
2		
3		

5. Design

Maximum (Mean) Duration of Followup	Number enrolled (total all arms)	Number randomized (total all arms)	Number of centers	Washout period, if XO (y/n/NA, duration) ⁷

6. Quality issues

Was Randomization Technique Reported? (y/n)	Was Allocation Concealment Method Reported? (y/n)	Claimed Intention to Treat Analysis in Methods (y/n)	Power calculation in Methods (y/n)	Verifiable† difference in reported results between text and table (y/n, what?) ⁷	
Blinded Patient (y/n/nd)	Blinded caregiver (y/n/nd)	Blinded Outcome Assessment (y/n/nd)	“Double blinded” (y/n)	“Single blinded” (y/n)	Other blinding (describe) ⁷

† A discrepancy you can point to, not just this “seems” wrong.

7. Outcomes. List all outcomes reported in the article (abstract, Results section, tables, figures). In round 1 of extraction complete only the first 5 columns (w/o ‡).

	Outcome Name ⁹ (Everyone)	In abstract? (y/n) (only non-English extractor) ¹⁰	Principal Timepoint for Outcome (only non-English extractor) ¹⁰	Dichot or Continuous? † (only non-English extractor) ¹⁰	Data/#s reported in text? (y/n) (only non-English extractor) ¹⁰	Chosen to extract results? (Y/N)‡ ¹⁰	Definition of Outcome (only if chosen to extract results)‡ ¹¹ (Everyone)
1							
2							
3							
4							

† Answer this for any outcome that is not obvious.

‡ After round 1, we will choose two outcomes (at 1 timepoint) for you to extract results. Outcome definitions are needed only for these 2 outcomes.

Tables 8&9: Extract results only for those outcomes listed in Table 7 under “Chosen to extract results”

8. Results (dichotomized outcomes). Include only reported data. Do not calculate any values.

Outcome	Intervention (intervention or control)	n Event	N Total	Unadjusted (reported)				Adjusted (reported)				
				Metric§	Result	95% CI#	P btw	Metric§	95% CI#	P btw	Adjusted for:	
	Tx											
	Cx											

§ RR, OR, HR, RD

Change to SD or SE, if necessary.

⁹ For analysis, used a maximum of 5 outcomes only.

¹⁰ Excluded from analysis (used only for organizational purposes).

¹¹ For two chosen outcomes only.

9. Results (continuous measures) Include only reported data. Do not calculate any values. If adjusted & unadjusted analyses reported, extract adjusted only.

Outcome	Intervention (intervention or control)	Unit	Baseline			Final			Change (Final – Baseline)			Net Δ /Difference* (Δ test – Δ control)*				
			N	Value	SD/SE*	N	Value	SD/SE*	Value	SD/SE/CI*	P	Value	SD/SE/CI*	P	Adjusted for:¶	
	Tx															
	Cx															

* Delete or correct the incorrect value/item. Replace with nd if necessary

¶ Complete only if analysis was adjusted (regardless of whether analysis was net change, difference of final values, or change from baseline). Otherwise leave blank.

10. Time. Only for those extracting from English translations

	A Lot	A Little	None
Extra time required to extract because of apparent poor translation:			

Appendix B. List of Articles Translated and Included

Chinese

Hu XY, Zhou YX, Xu SZ, et al. [Effects of probiotics on feeding intolerance in low birth weight premature infants]. [Chinese]. *Zhongguo Dangdai Erke Zazhi* 2010;12(9):693-5.

Li H, Dong L, Li Y, et al. [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]. *Chinese Journal of Lung Cancer* 2008;11(4):570-3.

Liu X, Liu D, Li J, et al. [Safety and efficacy of carbon dioxide insufflation during colonoscopy]. [Chinese]. *Zhong Nan da Xue Xue Bao* 2009;Yi(8):825-9.

Tang FZ, Liu YL, Wen FQ, et al. [Comparison of therapeutic effects in severe nocturia: gradual versus immediate drug withdrawal]. [Chinese]. *Zhongguo Dangdai Erke Zazhi* 2010;12(3):198-200.

Wang P, Yang J, Liu G, et al. [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/Journal of Chinese Integrative Medicine* 2010;8(7):636-40.

Xu JS, Yang JW, Gu MN, et al. [Effects of fentanyl on EC50 of ropivacaine for postoperative epidural analgesia after gynecological surgery]. [Chinese]. *Di Yi Junyi Daxue Xuebao* 2004;24(11):1326-7.

Xu XH, Chang YT, Li L, et al. [Effect of fructose-1,6-diphosphate on myocardial preservation during pulmonary operations]. [Chinese]. *Zhong Nan da Xue Xue Bao* 2008;Yi(10):966-9.

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/Journal of Chinese Integrative Medicine* 2010;8(3):231-7.

Yi JH, Li RR. [Influence of near-work and outdoor activities on myopia progression in school children]. [Chinese]. *Zhongguo Dangdai Erke Zazhi* 2011;13(1):32-5.

Zhang GQ, Ge L, Ding W, et al. [The value of portal vein chemotherapy after radical resection in delaying intrahepatic recurrence of stage II primary hepatocellular carcinoma]. [Chinese]. *Aizheng* 2008;27(12):1297-301.

French

Aubin M, Vezina L, Maziade J, et al. [Control of arterial hypertension: effectiveness of an intervention performed by family practitioners]. [French]. *Canadian Family Physician* 1994;40:1742-52.

Aydin A, Karadayi K, Aykan U, et al. [Effectiveness of topical ciclosporin A treatment after excision of primary pterygium and limbal conjunctival autograft]. [French][Erratum appears in J Fr Ophtalmol. 2010 Jun;33(6):435]. Journal Francais d Ophthalmologie 2008;31(7):699-704.

Baillargeon L, Drouin J, Desjardins L, et al. [The effects of Arnica Montana on blood coagulation. Randomized controlled trial]. [French][Erratum appears in Can Fam Physician 1994 Feb;40:225]. Canadian Family Physician 1993;39:2362-7.

Devogelaere T, Beresniak A, Raymaeckers A, et al. [Clinical study of Supranettes pads in the treatment of seasonal or perennial allergic conjunctivitis in children]. [French]. Journal Francais d Ophthalmologie 2006;29(6):593-8.

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]. Journal de Gynecologie, Obstetrique et Biologie de la Reproduction 2009;38(4):335-40.

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]. Journal Francais d Ophthalmologie 2008;31(1):51-5.

Gosselin P, Verreault R, Gaudreault C, et al. [Dietary treatment of mild to moderate hypercholesterolemia. Effectiveness of different interventions]. [French]. Canadian Family Physician 1996;42:2160-7.

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]. Gastroenterologie Clinique et Biologique 2000;24(5):495-500.

Polonovski JM, El MM. [Treatment of acute maxillary sinusitis in adults. Comparison of cefpodoxime-proxetil and amoxicillin-clavulanic acid]. [French]. Presse Medicale 2006;35(1:Pt 1):t-8.

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]. Gastroenterologie Clinique et Biologique 1997;21(12):924-8.

German

Bechdorf A, Pholmann B, Guttgemanns J, et al. Motivationsbehandlung für Patienten mit der Doppeldiagnose Psychose und Sucht **Ergebnisse einer randomisierten Studie**. Nervenarzt 2011;Epub ahead of print.

Birnbaum F, Schwartzkopff J, Bohringer D, et al. [Penetrating keratoplasty with intrastromal corneal ring. A prospective randomized study]. [German]. Ophthalmologie 2008;105(5):452-6.

Borner M, Burkle H, Trojan S, et al. [Intra-articular ketamine after arthroscopic knee surgery. Optimisation of postoperative analgesia]. [German]. Anaesthesist 2007;56(11):1120-7.

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(8):735-42.

Marx S, Cimniak U, Beckert R, et al. [Chronic prostatitis/chronic pelvic pain syndrome. Influence of osteopathic treatment - a randomized controlled study]. [German]. *Urologe (Ausg 2009;A)*.(11):1339-45.

Meybohm P, Hanss R, Bein B, et al. [Comparison of premedication regimes. A randomized, controlled trial]. [German]. *Anaesthesist* 2007;56(9):890-2.

Schnabel M, Vassiliou T, Schmidt T, et al. [Results of early mobilisation of acute whiplash injuries]. [German]. *Der Schmerz* 2002;16(1):15-21.

Stoffels I, Wolter TP, Sailer AM, et al. [The impact of silicone spray on scar formation. A single-center placebo-controlled double-blind trial]. [German]. *Hautarzt* 2010;61(4):332-8.

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(10):987-92.

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(11):1121-6.

Hebrew

Gimelfarb Y, Natan Z. [Risk factors for suicide attempts in dual diagnosis patients]. [Hebrew]. *Harefuah* 2009;148(6):355-8.

Haimov I, Vadas L. [Sleep in older adults: association between chronic insomnia and cognitive functioning]. [Hebrew]. *Harefuah* 2009;148(5):310-4.

Kugelman A, Anabussi S, Sharon N, et al. [The association between pertussis during infancy and childhood asthma]. [Hebrew]. *Harefuah* 2009;148(2):80-3.

Oksenberg A, Arons E, Greenberg-Dotan S, et al. [The significance of body posture on breathing abnormalities during sleep: data analysis of 2077 obstructive sleep apnea patients]. [Hebrew]. *Harefuah* 2009;148(5):304-9.

Oliven A, Tov N, Odeh M, et al. [Electrical stimulation of the genioglossus to improve pharyngeal patency in obstructive sleep apnea: comparison of results obtained during sleep and anesthesia]. [Hebrew]. *Harefuah* 2009;148(5):315-9.

Otto O, Peleg R, Press Y. [Streptococcal pharyngitis among children: comparison of attitudes between family physicians and pediatricians]. [Hebrew]. *Harefuah* 2009;148(8):511-4.

Perlitz Y, Gtezer-Soltzman S, Peleg A, et al. [Correlation of maternal serum and amniotic fluid leptin and insulin levels with neonatal birth weight]. [Hebrew]. *Harefuah* 2009;148(7):420-3.

Rosenberg E, Elkrinawi S, Goldbart A, et al. [Obstructive sleep apnea syndrome in young infants]. [Hebrew]. Harefuah 2009;148(5):295-9.

Italian

Berti M, Danelli G, Antonino FA, et al. 0.2% ropivacaine with or without sufentanil for patient-controlled epidural analgesia after anterior cruciate ligament repair. *Minerva Anestesiologica* 2005;71(3):93-100.

Borghi B, Laici C, Iuorio S, et al. [Epidural vs general anaesthesia]. [Italian]. *Minerva Anestesiologica* 2002;68(4):171-7.

Brizzi A, Greco F, Malvasi A, et al. Comparison of sequential combined spinal-epidural anesthesia and spinal anesthesia for cesarean section. *Minerva Anestesiologica* 2005;71(11):701-9.

Catania S, Gallo A. [Clinical efficacy and tolerability of short course therapy with cefaclor compared with long-term therapy for treatment of acute otitis media in children]. [Italian]. *Infezioni in Medicina* 2004;12(4):259-65.

Chisari G, Sanfilippo M, Reibaldi M. [Treatment of bacterial conjunctivitis with topical ciprofloxacin and norfloxacin: a comparative study]. [Italian]. *Infezioni in Medicina* 2003;11(1):25-30.

Cuomo G, Molinaro G, La MG, et al. [A comparison between the Simplified Disease Activity Index (SDAI) and the Disease Activity Score (DAS28) as measure of response to treatment in patients undergoing different therapeutic regimens]. [Italian]. *Reumatismo* 2006;58(1):22-5.

Guarda Nardini L, Oliviero F, Ramonda R, et al. [Influence of intra-articular injections of sodium hyaluronate on clinical features and synovial fluid nitric oxide levels of temporomandibular osteoarthritis]. [Italian]. *Reumatismo* 2004;56(4):272-7.

Lo Martire N, Savastano S, Rossini L, et al. [Topical anesthesia for cataract surgery with phacoemulsification: lidocaine 2% vs ropivacaine 1%. Preliminary results]. [Italian]. *Minerva Anestesiologica* 2002;68(6):529-35.

Nieddu ME, Menza L, Baldi F, et al. [Efficacy of Cellfood's therapy (deutrosulfazyme) in fibromyalgia]. [Italian]. *Reumatismo* 2007;59(4):316-21.

Pasero GP, Di MO. [Analgesic dose range finding of lornoxicam compared to diclofenac. Crossover double blind study in rheumatoid arthritis]. [Italian]. *Reumatismo* 2002;54(3):238-42.

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 - Japanese Journal of Clinical Hematology* 2010;51(5):315-9.

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 [Japanese Journal of Cancer & Chemotherapy] 2009;36(1):77-82.

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]. Journal of Oleo Science 2008;57(1):23-33.

Miura H, Takahashi Y, Kitabatake Y. [Influence of group training on pulse wave velocity in elderly women]. [Japanese]. Nippon Koshu Eisei Zasshi - Japanese Journal of Public Health 2010;57(4):271-8.

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 - Journal of the Pharmaceutical Society of Japan 2004;124(12):989-95.

Satou Y, Kanda J, Okumura M, et al. [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(4):117-21.

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 - Japanese Journal of Clinical Hematology 2009;50(12):1692-9.

Sekine Y, Takai Y, Nishii O, et al. [Establishment of an optimum bowel preparation method before gynecologic laparoscopic surgery]. [Japanese]. Yakugaku Zasshi - Journal of the Pharmaceutical Society of Japan 2001;121(8):637-45.

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]. Japanese Journal of Medical Mycology 2001;42(4):223-8.

Takahashi M, Araki A, Ito H. [Development of a new method for simple dietary education in elderly individuals with diabetes mellitus]. [Japanese]. Nippon Ronen Igakkai Zasshi - Japanese Journal of Geriatrics 2002;39(5):527-32.

Korean

Boo GB, Oh JC, Lee BJ, et al. [The effect of proton pump inhibitor on healing of post-esophageal variceal ligation ulcers]. [Korean]. Korean Journal of Gastroenterology/Taehan Sohewagi Hakhoe Chi 2008;51(4):232-40.

Cho SB, Park KJ, Lee JS, et al. [Comparison of terlipressin and octreotide with variceal ligation for controlling acute esophageal variceal bleeding--a randomized prospective study]. [Korean]. Korean Journal of Hepatology 2006;12(3):385-93.

Choi WH, Park DI, Oh SJ, et al. [Effectiveness of 10 day-sequential therapy for Helicobacter pylori eradication in Korea]. [Korean]. Korean Journal of Gastroenterology/Taehan Sohwagi Hakhoe Chi 2008;51(5):280-4.

Gwak JH, Kim JY, Kim HJ, et al. [The Effect of Isoflavone and Gamma-linolenic Acid Supplementation on Serum Lipids and Menopausal Symptoms in Postmenopausal Women]. [Korean]. Korean J Nutr 2010;43(2):123-31.

Jeong HY, Lee BS, Sung JK, et al. [A randomized, prospective, comparative, multicenter study of rabeprazole and ranitidine in the treatment of reflux esophagitis]. [Korean]. Korean Journal of Gastroenterology/Taehan Sohwagi Hakhoe Chi 2006;47(1):15-21.

Kim JS, Kim HJ, Woo YH, et al. [Effects on changes in femoral vein blood flow velocity with the use of lower extremity compression for critical patients with brain injury]. [Korean]. Journal of Korean Academy of Nursing 2009;39(2):288-97.

Kim YG, Moon JT, Lee KM, et al. [The effects of probiotics on symptoms of irritable bowel syndrome]. [Korean]. Korean Journal of Gastroenterology/Taehan Sohwagi Hakhoe Chi 2006;47(6):413-9.

Lee SS, Yoon H. [A comparison of the effect of lidocaine or sodium bicarbonate mixed with rocuronium on withdrawal movement, mean arterial pressure and heart rate during rocuronium injection]. [Korean]. Journal of Korean Academy of Nursing 2009;39(2):270-8.

Park JY, Kim JY, Lee SP, et al. [The Effect of Green Coffee Bean Extract Supplementation on Body Fat Reduction in Overweight/Obese Women]. [Korean]. Korean J Nutr 2010;43(4):374-81.

Seo YJ, Yoon H. [The effects of preemptive analgesia of morphine and ketorolac on postoperative pain, cortisol, O₂ saturation and heart rate]. [Korean]. Journal of Korean Academy of Nursing 2008;38(5):720-9.

Portuguese

Amorim MM, Lippo LA, Costa AA, et al. [Transdermal nitroglycerin versus oral nifedipine administration for tocolysis: a randomized clinical trial]. [Portuguese]. Revista Brasileira de Ginecologia e Obstetricia 2009;31(11):552-8.

Camargo MA, Lopes LR, Grangeia TA, et al. [Use of corticosteroids after esophageal dilations on patients with corrosive stenosis: prospective, randomized and double-blind study]. [Portuguese]. Revista Da Associacao Medica Brasileira 2003;49(3):286-92.

Carramao S, Auge AP, Pacetta AM, et al. [A randomized comparison of two vaginal procedures for the treatment of uterine prolapse using polypropylene mesh: hysteropexy versus hysterectomy]. [Portuguese]. Revista do Colegio Brasileiro de Cirurgioes 2009;36(1):65-72.

de Arruda LH, Kodani V, Bastos FA, et al. [A prospective, randomized, open and comparative study to evaluate the safety and efficacy of blue light treatment versus a topical benzoyl peroxide 5% formulation in patients with acne grade II and III]. [Portuguese]. *Anais Brasileiros de Dermatologia* 2009;84(5):463-8.

Machado AF, Pedreira ML, Chaud MN. [Prospective, randomized and controlled trial on the dwell time of peripheral intravenous catheters in children, according to three dressing regimens]. [Portuguese]. *Revista Latino-Americana de Enfermagem* 2005;13(3):291-8.

Magacho L, Reis R, Pignini MA, et al. [2% ibopamine vs. water-drinking test as a provocative test for glaucoma]. [Portuguese]. *Arquivos Brasileiros de Oftalmologia* 2008;71(4):499-503.

Muller KR, Bonamigo RR, Crestani TA, et al. [Evaluation of patients' learning about the ABCD rule: A randomized study in southern Brazil]. [Portuguese]. *Anais Brasileiros de Dermatologia* 2009;84(6):593-8.

Pereira PP, Oliveira AL, Cabar FR, et al. [Comparative study of manual vacuum aspiration and uterine curettage for treatment of abortion]. [Portuguese]. *Revista Da Associacao Medica Brasileira* 2006;52(5):304-7.

Santos FM, Rodrigues RG, Trindade-Filho EM. [Physical exercise versus exercise program using electrical stimulation devices for home use]. [Portuguese]. *Revista de Saude Publica* 2008;42(1):117-22.

Simao AN, Godeny P, Lozovoy MA, et al. [Effect of n-3 fatty acids in glycemic and lipid profiles, oxidative stress and total antioxidant capacity in patients with the metabolic syndrome]. [Portuguese]. *Arquivos Brasileiros de Endocrinologia e Metabologia* 2010;54(5):463-9.

Spanish

Bonetto G, Salvatico E, Varela N, et al. [Pain prevention in term neonates: randomized trial for three methods]. [Spanish]. *Archivos Argentinos de Pediatria* 2008;106(5):392-6.

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]. *Archivos Argentinos de Pediatria* 2010;108(3):201-8.

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]. *Nutricion Hospitalaria* 2010;25(1):53-9.

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]. *Nutricion Hospitalaria* 2010;25(4):606-12.

Gomez-Garcia A, Hernandez-Salazar E, Gonzalez-Ortiz M, et al. [Effect of oral zinc administration on insulin sensitivity, leptin and androgens in obese males]. [Spanish]. *Revista Medica de Chile* 2006;134(3):279-84.

Lopez-De-Blanc SA, Salati-De-Mugnolo N, Femopase FL, et al. Antifungal topical therapy in oral chronic candidosis. A comparative study. *Medicina Oral* 2002;7(4):260-70.

Martinez Gonzalez JM, Benito PB, Fernandez CF, et al. A comparative study of direct mandibular nerve block and the Akinosi technique. *Medicina Oral* 2003;8(2):143-9.

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 (Asturias, Spain)* 2005;16(1):5-12.

Rodriguez Martin C, Castano Sanchez C, Garcia Ortiz L, et al. [Efficacy of an educational intervention group on changes in lifestyles in hypertensive patients in primary care: a randomized clinical trial]. [Spanish]. *Revista Espanola de Salud Publica* 2009;83(3):441-52.

Vasquez AM, Sanin F, Alvarez LG, et al. [Therapeutic efficacy of a regimen of artesunate-mefloquine-primaquine treatment for Plasmodium falciparum malaria and treatment effects on gametocytic development]. [Spanish]. *Biomedica* 2009;29(2):307-19.

English

Artinian NT, Flack JM, Nordstrom CK, et al. Effects of nurse-managed telemonitoring on blood pressure at 12-month follow-up among urban African Americans. *Nurs Res* 2007;56(5):312-22.

Binanay C, Califf RM, Hasselblad V, et al. Evaluation study of congestive heart failure and pulmonary artery catheterization effectiveness: the ESCAPE trial. *JAMA* 2005;294(13):1625-33.

Bjorkman M, Sorva A, Risteli J, et al. Vitamin D supplementation has minor effects on parathyroid hormone and bone turnover markers in vitamin D-deficient bedridden older patients. *Age Ageing* 2008;37(1):25-31.

Calo L, Lamberti F, Loricchio ML, et al. Left atrial ablation versus biatrial ablation for persistent and permanent atrial fibrillation: a prospective and randomized study. *J Am Coll Cardiol* 2006;47(12):2504-12.

Domagk D, Menzel J, Seidel M, et al. Endoluminal gastroplasty (EndoCinch) versus endoscopic polymer implantation (Enteryx) for treatment of gastroesophageal reflux disease: 6-month results of a prospective, randomized trial. *Am J Gastroenterol* 2006;101(3):422-30.

Gagnadoux F, Fleury B, Vielle B, et al. Titrated mandibular advancement versus positive airway pressure for sleep apnoea. *Eur Respir J* 2009;34(4):914-20.

Meador K, Loring D, Nichols M, et al. Preliminary findings of high-dose thiamine in dementia of Alzheimer's type. *J Geriatr Psychiatry Neurol* 1993;6(4):222-9.

Noel-Weiss J, Rupp A, Cragg B, et al. Randomized controlled trial to determine effects of prenatal breastfeeding workshop on maternal breastfeeding self-efficacy and breastfeeding duration. *J Obstet Gynecol Neonatal Nurs* 2006;35(5):616-24.

Schiele F, Meneveau N, Vuilleminot A, et al. Impact of intravascular ultrasound guidance in stent deployment on 6-month restenosis rate: a multicenter, randomized study comparing two strategies--with and without intravascular ultrasound guidance. RESIST Study Group. REStenosis after Ivus guided STenting. J Am Coll Cardiol 1998;32(2):320-8.

Webster J, Marshall F, Abdalla M, et al. Randomised comparison of percutaneous angioplasty vs continued medical therapy for hypertensive patients with atheromatous renal artery stenosis. Scottish and Newcastle Renal Artery Stenosis Collaborative Group. J Hum Hypertens 1998;12(5):329-35.

Appendix C. Examples of Poorly Translated Articles

This appendix includes examples of difficult to interpret, incomplete, mangled, or otherwise difficult translations. The translations are organized by language. The appendix contains only de-identified fragments.

Chinese

“Gao oxaliplatin (GO)”

“advanced NSCLC 51 例”

“party is ginseng, astragalus, stabbed five tomato, cantharidin, the main components of ginseng saponin, astragalus saponin, thorn Five tomato polysaccharide to a cantharidin,”

“1.4.2 KPS score of [2], after treatment than before treatment score ≥ 10 points were added to improve, increase or decrease <10 points for the stability, reduce by ≥ 10 points for the decline.”

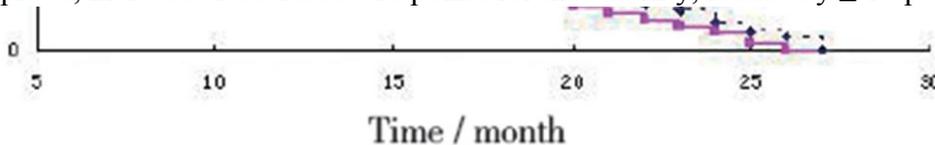


图 1 治疗组与对照组Kaplan-Meier生存曲线

Fig 1 Kaplan-Meier survival curves of treatment and control groups

Figure 1. Untranslated x-axis legend from a Chinese article.

“Evaluate the purpose of carbon dioxide, \$ [F"% colonoscopy is the safety and effectiveness of "Methods to% & \$ # patients were randomly divided into [F" Group \$ \$ j6] &% \$ \$, and air group j6] 5%! were injected”

“HI Materials and Methods

6 / 6! Study

The "# #" years \$ month" to "# #" # 6 in the period from 66 months in the”

“By ZUZZ6% / # statistical package was used for statistical analysis”

“!; A,, anesthesia

Patients were tested before surgery A \$, 17: intramuscular injection of atropine \$; #, 13, and stability! \$, 13 # into the operating room after the d! - "or d"-A line of epidural space puncture”

French

“SUMMARY

(ETFE study under fGnlt of Eshi li Randomized double-blind study plan AOLS, deyait yirfler a sl miclklllMllf honaiopathique, Anka Montana eu le sigli6cativement time scignement (SllllpIate II) and its effects dialre sw diffirents Hngulne coagulation tests. II was that prodllt Inll .. on various parameters of coagulation Ia Hngulne among yolontalres ss Gll cows minutes Sliva ... Sin administration.”

“Thirty minutes after taking the second dose tube, the technician again efTectuait bleeding time and collect tubes of blood.”

SUJET	PERIODE 1	
	T ₀	T ₃₀
9	5,0	6,5
10	8,0	8,0
11	5,0	5,5
12	5,5	6,5
13	5,0	4,0
14	6,3	5,3
15	6,0	7,5
16	5,0	5,8
17	7,5	6,8
MOYENNE	5,92	6,19

Figure 2. A table with untranslated text (“sujet”, “periode”, “moyenne”) from a French article.

[Added by Research Assistant:

Groupe = group

Moyenne = average

Periode = period

* T: Time of initial bleeding

** T3: Bleeding time 30 minutes after treatment

*** Y, and Y2: T3-To]

“Acknowledgements This study ae subsidized by the Fonds de la recherche en santé du Québec.”

“Mari6 (%)”

“C-1, DL (mmol / L)”

“| *. gloomy subjects in each group.”

sinusitis in adults
oxetil-

OBJECTIVE: The objective of this study was to evaluate the efficacy, safety, and tolerance of cefpodoxime proxetil (200 mg) plus clavulanic acid (1 g/125 mg) in the treatment of acute maxillary sinusitis in adults. In this prospective, multicenter, randomized, open, 512 adult patients with maxillary sinusitis unilateral according to criteria established by the Canadian Council of Health Products (AFSSAPS) were treated with cefpodoxime proxetil (200 mg) plus clavulanic acid (1 g/125 mg) for 5 days. The clinical success rate between the two groups was not significantly different (92.3% (215/233) in the cefpodoxime proxetil group and 93.6% (204/218) in the cefpodoxime proxetil plus clavulanic acid group).

Figure 3. Translated text where columns of text overlap heavily so that large portions of the text are unreadable.

“Subjects Tanvahtiyim. Subjects with Ednge”that my position had fewer sleep breathing disorders”

“Subjects were not Tanvahtiyim many breathing disorders are followed Sacheveat back and after Sacheveat Party, for continuous positive airway pressure (CPAP (“

“Table 2 also concentrated nocturnal Polisumanugerfeim data subjects' Suffering Bdnge”

“This work demonstrated that% 53.8 of 2077 patients”

“the criteria K to N. Y. Y. Y. Y. F from a D from S. F. J. R. C. K. S. J. and S. Y. D and C to D to Y. A's in H and N D to K and Z to A.”

13.	רפאפן	*27	57.4	52	100	0.0001
14.	מוקסיפן	20	34.2	0	0	
15.	באיזה גיל ובשנים לדעתך עולה הסבירות שמדובר בדלקת לוע סטרפטוקוקית?					
16.	14-3	45	94	50	96	NS

Figure 7. Untranslated Hebrew in a column of data within a table (rows 13 to 16).

[Research Assistant added:

13. Rafapne

14. Muxipne

15. At what age) in (do you think the more likely it is pharyngitis Straftokoakit?

16.14-3]

Italian



Figure 8. Italian text with Google translate balloon: “Contribute a better translation.”

nausea, vomiting, itching than those who had received sufentanil with a longer time than necessary to the duct, while Palm *et al.* 2 have showed that 0.75 $\mu\text{g} / \text{ml}$ sufentanil decreased by almost 30% as the lowest concentration of anesthetic necessary for a local delivery of effective analgesia in 50% of pregnant women. Demonstrated the addition of 5, 10 and 15 mg of sufentanil in 12 ml of ropivacaine 0.2% and sufentanil showed that the improvement spaces L3-L4 or L4-L5 using a technique need-

0.2% for epidural analgesia by patient-controlled (PCEA) on the quality of postoperative pain in patients undergoing chronic ligament reconstruction of the anterior cruciate ligament. Methods: We enrolled 20 patients ASA physical status I and II are candidates for surgery to rebuild the anterior cruciate ligament. It was a self-made combined spinal-epidural anesthesia at the level of the interspaces L3-L4 or L4-L5 using a technique need-

Figure 9. Translated text where columns of text overlap heavily so that large portions of the text are unreadable.

Japanese

“Ⓢ The laboratory values of WBC $\geq 3,000 / \text{Mm}^3$, $i \geq$ platelets Number $10 \times 10^4 / \text{mm}^3$, $6 \geq$ total protein.Og / dL (A / G ≥ 1.0), AST, ALT $100 \leq, \leq$ serum creatinine 1.5 mg / d Shino patients meet the criteria.

Registration of cases, after the eligibility check at the Central Registration System to the intestines or rectum before forming another layer, were provisionally registered. Submitted 10 months after surgery - 12 "10 months a progress report on the treatment" to exclude based on the following cases, the remaining cases were registered for this."

"total dose prescribed period of 80 or more ⑤% doing therapy."

"1. Target

慶應義塾大学 period from March 1997 until August 1996"

"6. See observation day and observe

1) Sun observation

Hazime Higai administration was required to observe the day and 4 weeks after 2 weeks."

"8. Stop loss criteria"

Figure 10. Translated text where columns of text overlap heavily so that large portions of the text are unreadable.

Research Assistant's table:
Medication status in Table 3

Actual duration of treatment	Treated group 1 year (n=108)	Three treatment groups in (n=113)
Less than 10 months		
10 to 12 months		

Which had to be matched with:

表 3 投薬状況

実投与期間	1年投与群 (n=108)	3年投与群 (n=113)
10か月未満	1 (0.9%)	
10~12か月	46 (42.6%)	

Figure 11. Untranslated Japanese within a table.

Korean

"four with a flexible over-tube was yongha."

"proximal wibunmunbueseo"

"The number of patients the PPI group, 25 patients were 22 patients in the control group this average age of four the two groups, there was no significant difference in gender ratio."

"High kicks

Last EVL for acute bleeding caused by..."

"2. Endoscopy chiyuyul"

"cure rates for my PD ransopeurajol 30 mg 75-93%, rabepeurajol 20 mg 76-92%"

Portuguese

“There may perfuração”

“Between 20% and 40% of patients with grade II and III lesions will develop estenose”

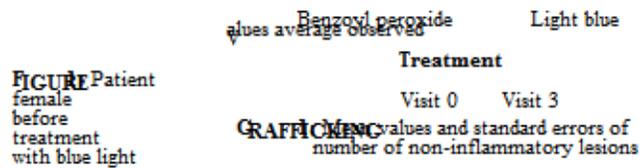


Figure 12. Portion of a translated table with overlapping text and poor spacing of various words making the meaning unintelligible.

Spanish

The age of volunteers was 21.8 ± 2.8 and 25.1 ± 4.5 years for the GZn and GC, respectively mind, ($p = 0.318$). Both groups had a BMI similar (30.7 ± 2.6 vs. 30.5 ± 3.9 , $p = 0.910$ for GZn and GC, respectively).

Figure 13. Overlapping text.