Draft Methods Report

Number XX

Assessing the Accuracy of Google Translate for the Purpose of Translating non-English Language Trials for Data Extraction

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

This information does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services determination or policy.

Contract No. [redacted]

Prepared by:
[redacted]

Investigators
[redacted]

Task Order Officer
Kim Marie Wittenberg, MA

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.

Suggested Citation: pdg. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

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

AHRQ Contact

Kim Marie Wittenberg, M.A.
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850

This report is based on research conducted by the [redacted] Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. [redacted]). 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 clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the 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.

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 e-mail to epc@ahrq.hhs.gov.

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

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

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

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

Contents

Abstract v
Introduction 1
Aims 2
Methods 3
Study selection 3
Translation 3
Data extraction 5
Results 7
Study selection 7
Article translation 7
Data extraction from translated articles 8
Comparison of translated with original articles 9
Discussion 21
Abbreviations 22
References 23

Tables

Table 1. Percentage of studies from Medline in various languages 1
Table 2. Translation time, by language 8
Table 3. Agreement across all items, by language 10
Table 4. Percent of items for with there was agreement, by language and data extraction form section 19
Table 5. Items with at least 80 percent agreement, by language and data extraction form section 20

Figures

Figure 1. Histograms of the percent agreement for all items, by language 11
Figure 2. Histograms of the percent agreement for eligibility criteria items, by language 12
Figure 3. Histograms of the percent agreement for descriptions of treatment and control items, by language 13
Figure 4. Histograms of the percent agreement for study characteristics items, by language 14
Figure 5. Histograms of the percent agreement for study methodology items, by language 15
Figure 6. Histograms of the percent agreement for descriptions of outcomes items, by language 16
Figure 7. Histograms of the percent agreement for categorical results items, by language 17
Figure 8. Histograms of the percent agreement for continuous results items, by language 18

Appendices

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

Appendices cited in this report are available at http://ahrq.gov/XXXXX **[AHRQ will supply the rest of the information for this statement.]**

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 8 languages (Chinese, French, German, Italian, Japanese, Korean, Portuguese, and Spanish) and 8 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 1 of 5 current EPC researchers who did not speak the given language. These 5 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%) to about 1 hour. Assessment by the English language data extractors indicated that "a little" extra time was required for 40 articles (45%) and "a lot" for 42 (48%). 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% of the items, compared with 80% in English articles. Spanish, Hebrew, and Chinese had the lowest agreement (30%, 24%, and 8%, 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.

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

A Medline search of all publications from 2000 to February 3, 2011 found that of 6,574,939 citations, 90 percent are published in English. Table 1 shows the number and frequency of publications in other languages with at least 1 percent frequency. In addition, the [redacted] EPC has recently had to account for articles in Portuguese, Italian, Dutch, and Hebrew.

Table 1. Percentage of studies from Medline in various languages

Language	N	%
20$.ed (Total)	6,574,939	100%
English	5,926,763	90%
Chinese	109,658	1.7%
French	97,752	1.5%
German	88,191	1.3%
Japanese	73,657	1.1%
Russian	71,583	1.1%
Spanish	71,281	1.1%

In general, the [redacted] 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.

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 from 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 in the lay press. If the extractions 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.{Sheppard, 2011 129 /id}

Aims

We conducted a pilot study to formally evaluate the accuracy of Google Translate for the purposes of data extraction of non-English language articles. We performed simultaneous limited data extraction of a random selection of recently published non-English language publications and their Google translations.

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 [redacted] EPC, we included articles in the following 9 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 [redacted] 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). 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 (that we could translate 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). As 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. Download Chrome. (This was done under the theory that Google Translate would be most functional in the Google browser, Chrome.)
2.
    a. If you're working from a pdf
        i. Go to http://translate.google.com/#
        ii. Under the large text box, in light blue, click "translate a document"
        iii. Browse to the relevant pdf/htm.
        iv. Pick the From language.
        v. Click Translate
    b. You may get a pop-up at the top of the page asking if you would like to translate. Translate.
    c. This approach is most likely to work with recent articles from journals that produce easily readable pdf files in Latin alphabets.
3. Save the translation as an html file.
4. 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)
5. 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.
6. 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
7. 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 [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 differences 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 technique, use of blinding, allocation concealment, statistical methods, 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.

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."

[*] Initials in brackets refer to the study investigators (authors) or acknowledged colleagues.

Data extraction comparison

For each study, a single researcher [EB], 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) was 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 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 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%), two in Chinese (2/12, 17%), three in Korean (3/13, 23%), four in Italian (4/14, 29%), nine in Hebrew (9/17, 53%), 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).

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

The length of time required to translate articles ranged from seconds (51 of 88 articles, 58%) 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	secs	secs	secs	secs	secs	secs	30 min	45 min	60 min
German	secs	secs	secs	secs	secs	secs	secs	secs	20 min	20 min
Italian	secs	secs	secs	secs	secs	secs	secs	secs	secs	1 min
Portuguese	secs	secs	secs	secs	secs	secs	secs	secs	secs	45 min
Spanish	secs	secs	secs	secs	secs	secs	secs	secs	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	secs	secs	secs	secs	secs	secs	20 min	45 min	60 min
Korean	10 min	10 min	10 min	10 min	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 = minutes; secs = "seconds"

* 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 (artifactual breaks not at the end of sentences) in the Asian language articles to allow meaningful translation. Translation of tables were frequently very time-consuming as they 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 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%) 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%) "a little" extra time was required and for 42 (48%) "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% "a lot"), German (70%), Japanese (70%), Korean (50%), Hebrew (50%). 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 ≧ 3,000 / mm^3, i ≧ platelets Number $10xlO^4/mm^3$, 6 ≧ total protein.O g/dL (A/G ≧ 1.0), AST, ALT 100 ≦, ≦ 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. 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 ≥80% Agreement	P*	Articles with ≥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)		53/66 (80%)		8/10 (80%)	

*P value versus English

Tables 4 and 5 display the absolute agreement (Table 4) and the percent of items with high agreement (Table 5) 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.

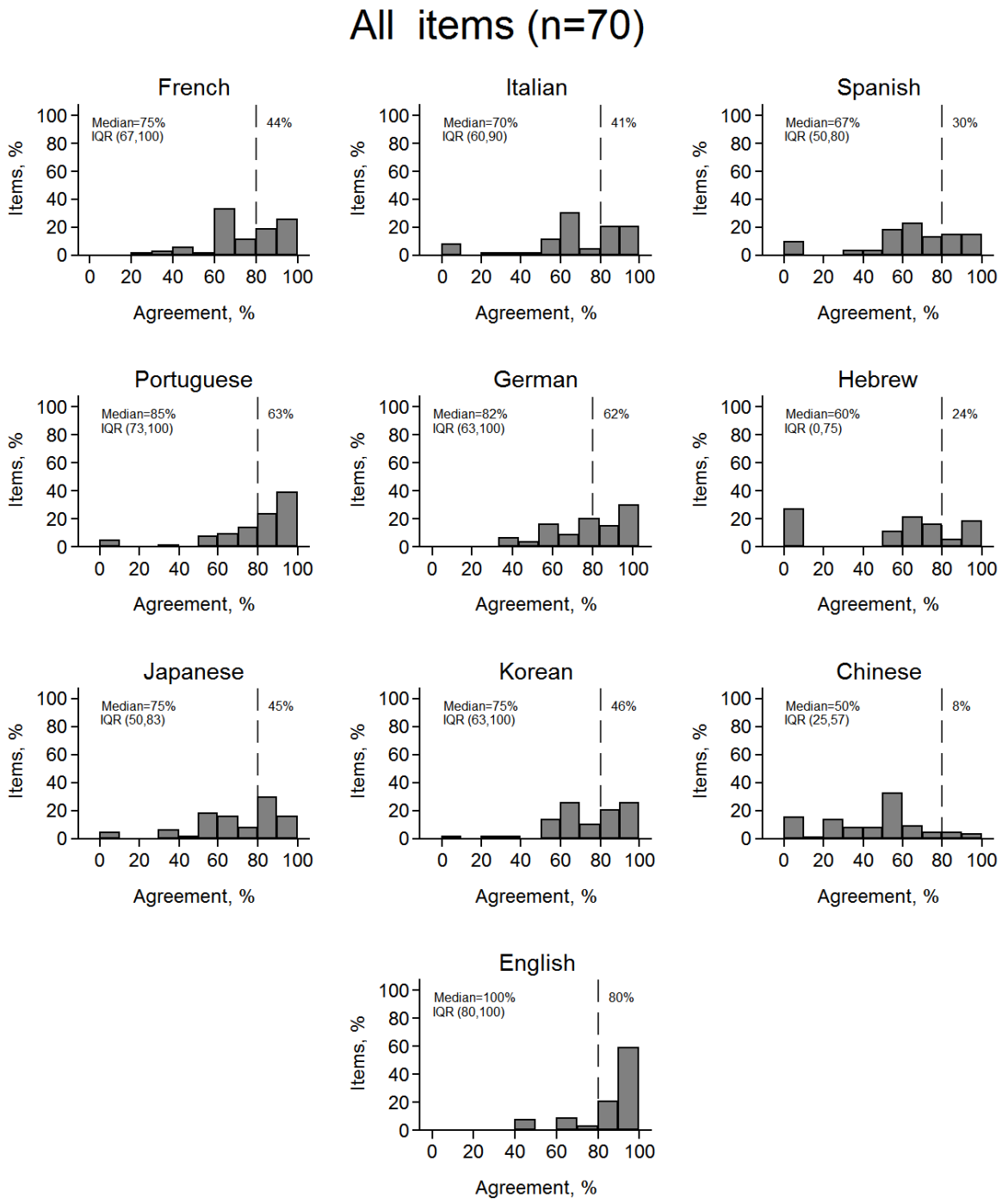

Figure 1. Histograms of the percent agreement for all items, by language. There were a maximum of 70 items per study. All histograms are drawn on the same scale, with 10 percentage point wide bars. For each histogram (language), the median and interquartile range (IQR) percent agreement across the 10 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% agreement (indicated by the vertical dashed line).

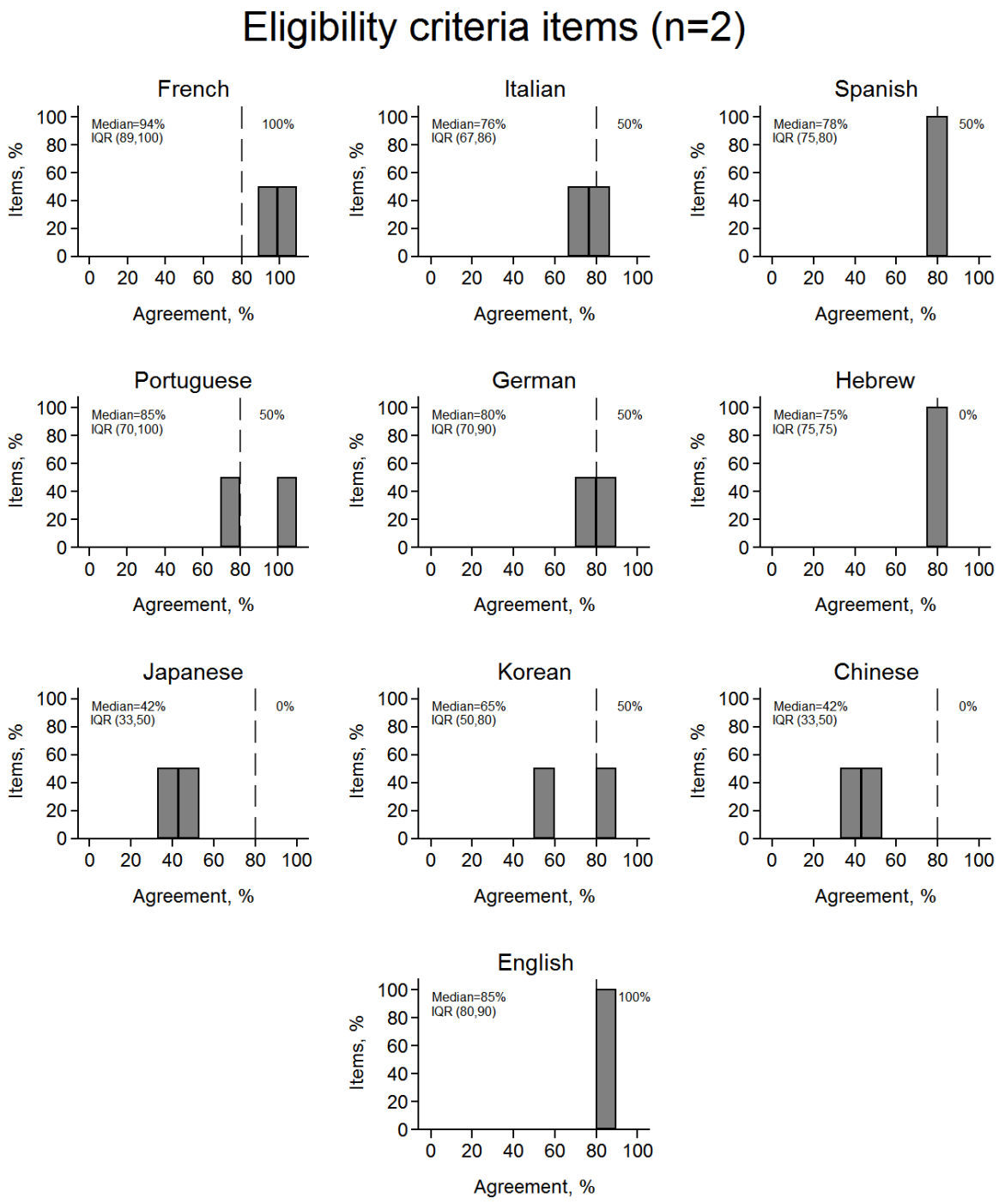

Figure 2. Histograms of the percent agreement for eligibility criteria items, by language. There were a maximum of 2 such items per study. See legend to Figure 1.

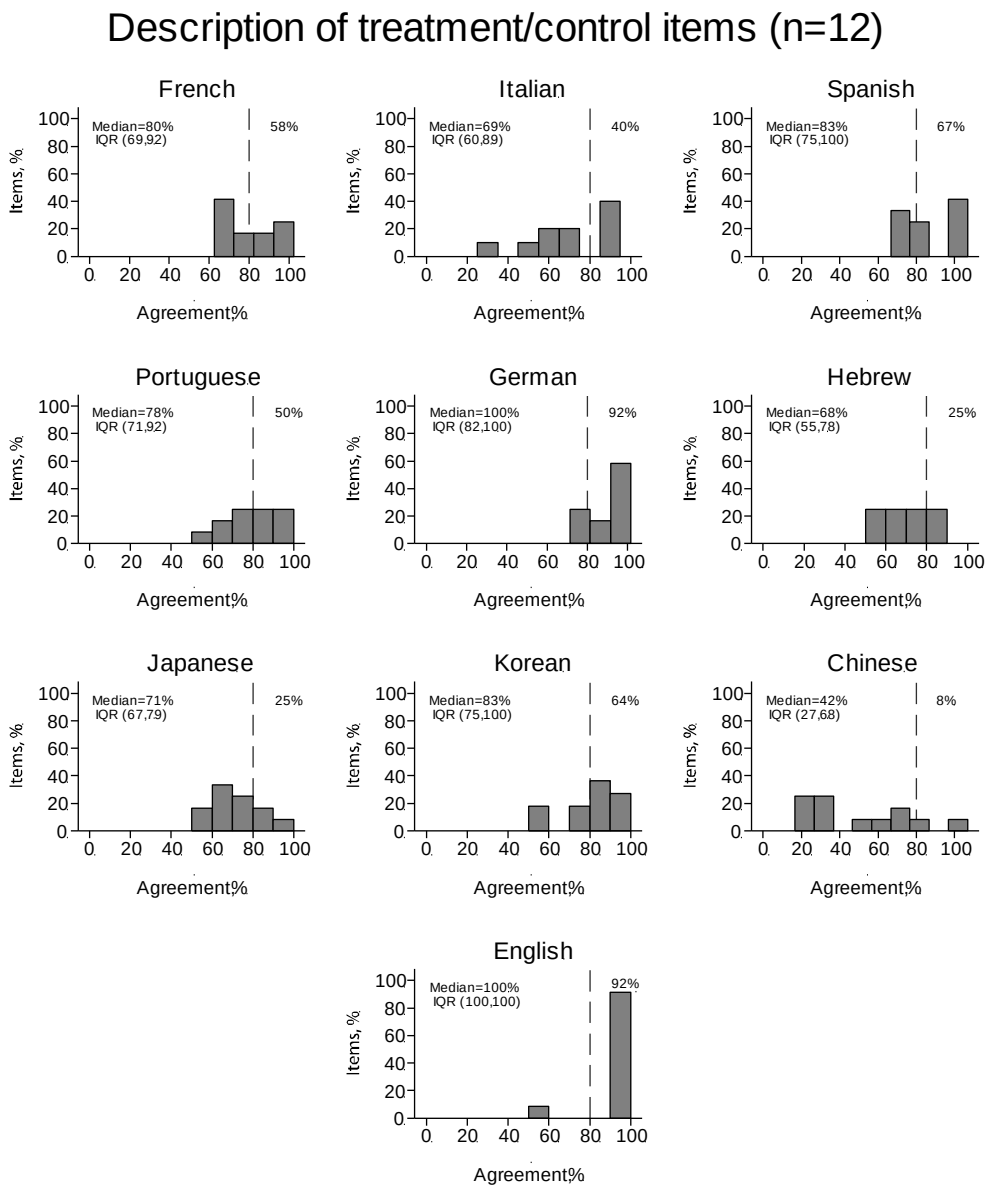

Figure 3. Histograms of the percent agreement for descriptions of treatment and control items, by language. There were a maximum of 12 such items per study. See legend to Figure 1.

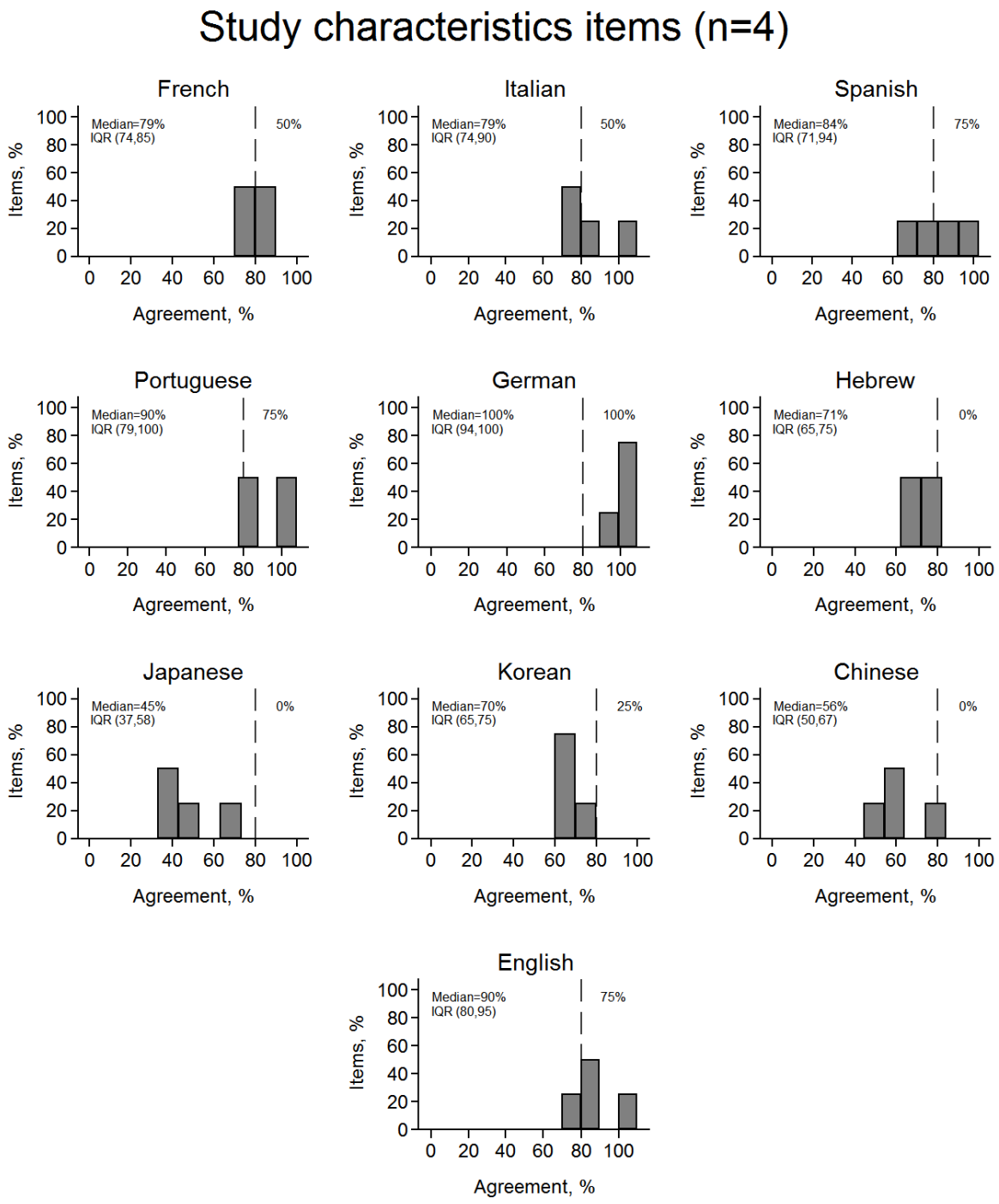

Figure 4. Histograms of the percent agreement for study characteristics items, by language. There were a maximum of 4 such items per study. See legend to Figure 1.

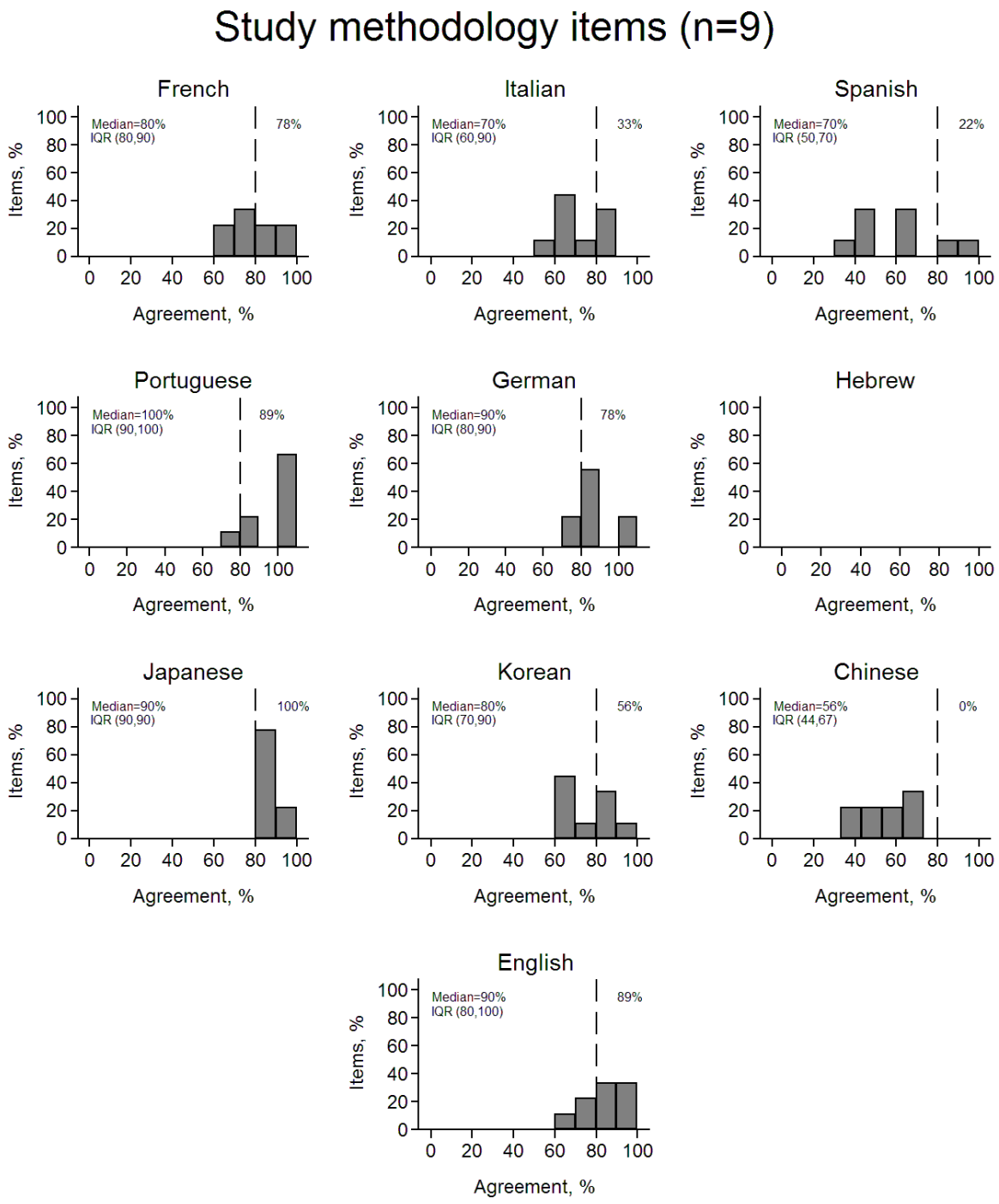

Figure 5. Histograms of the percent agreement for study methodology items, by language. There were a maximum of 9 such items per study. See legend to Figure 1.

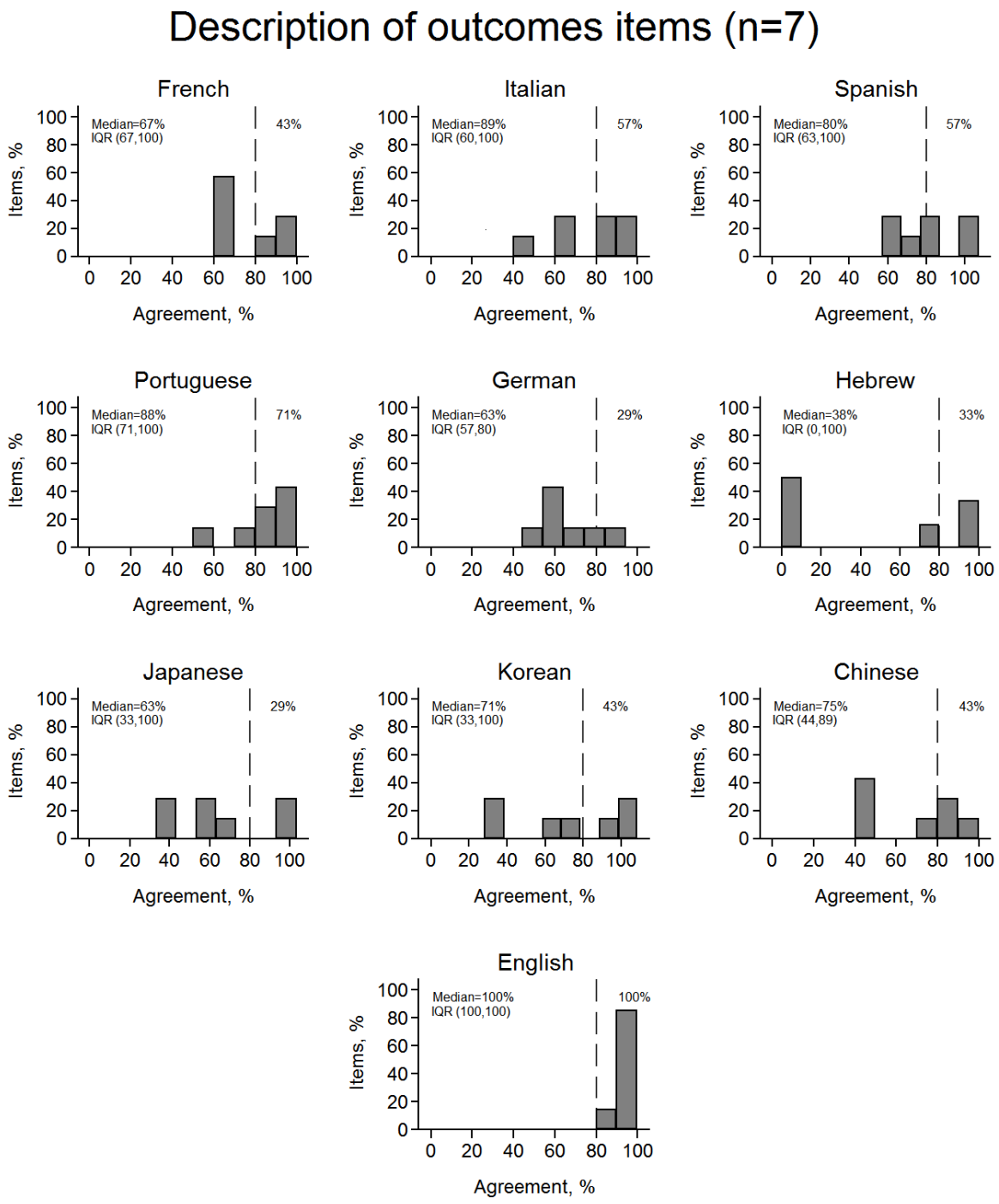

Figure 6. Histograms of the percent agreement for descriptions of outcomes items, by language. There were a maximum of 7 such items per study. See legend to Figure 1.

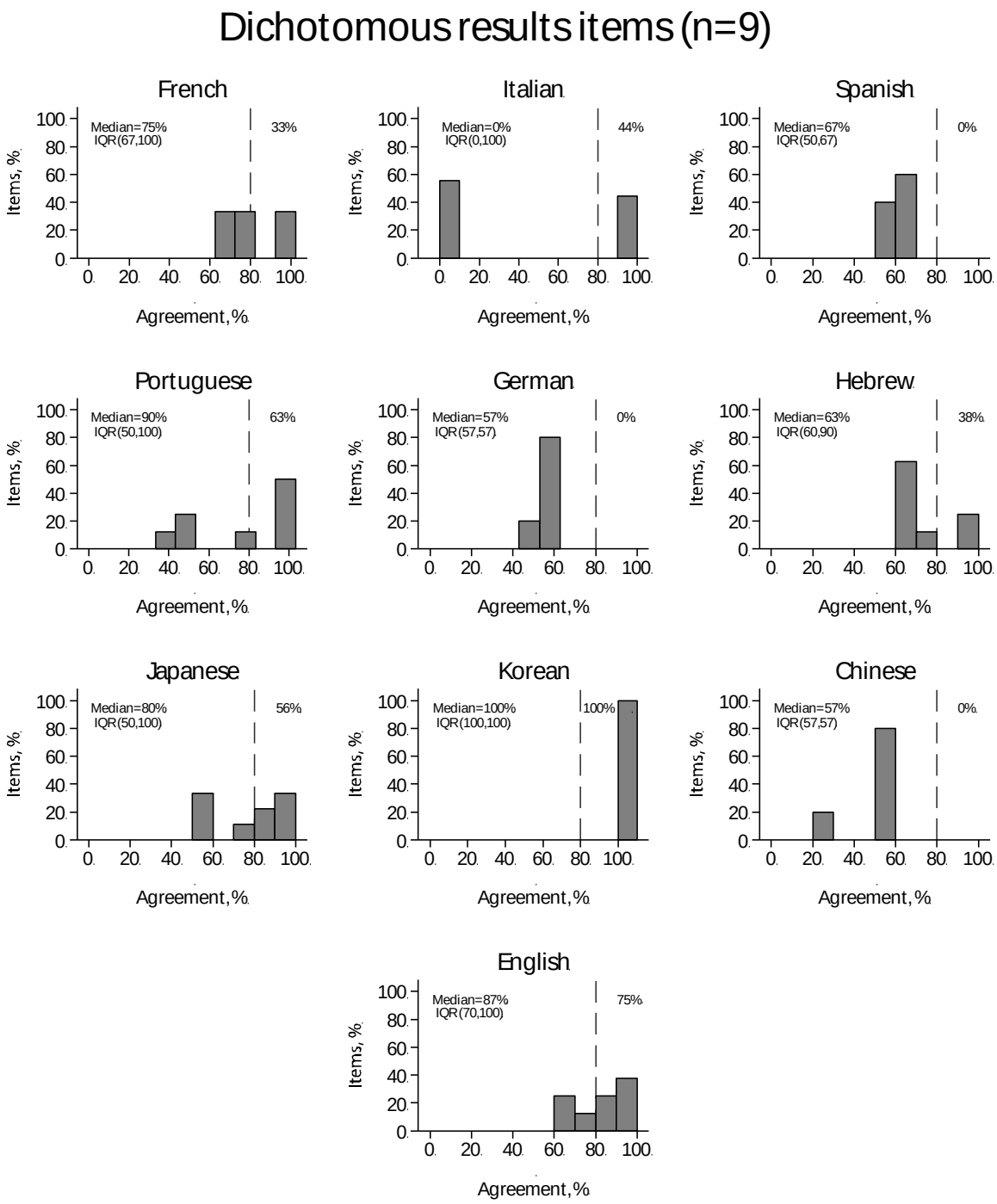

Figure 7. Histograms of the percent agreement for categorical results items, by language. There were a maximum of 9 such items per study. See legend to Figure 1.

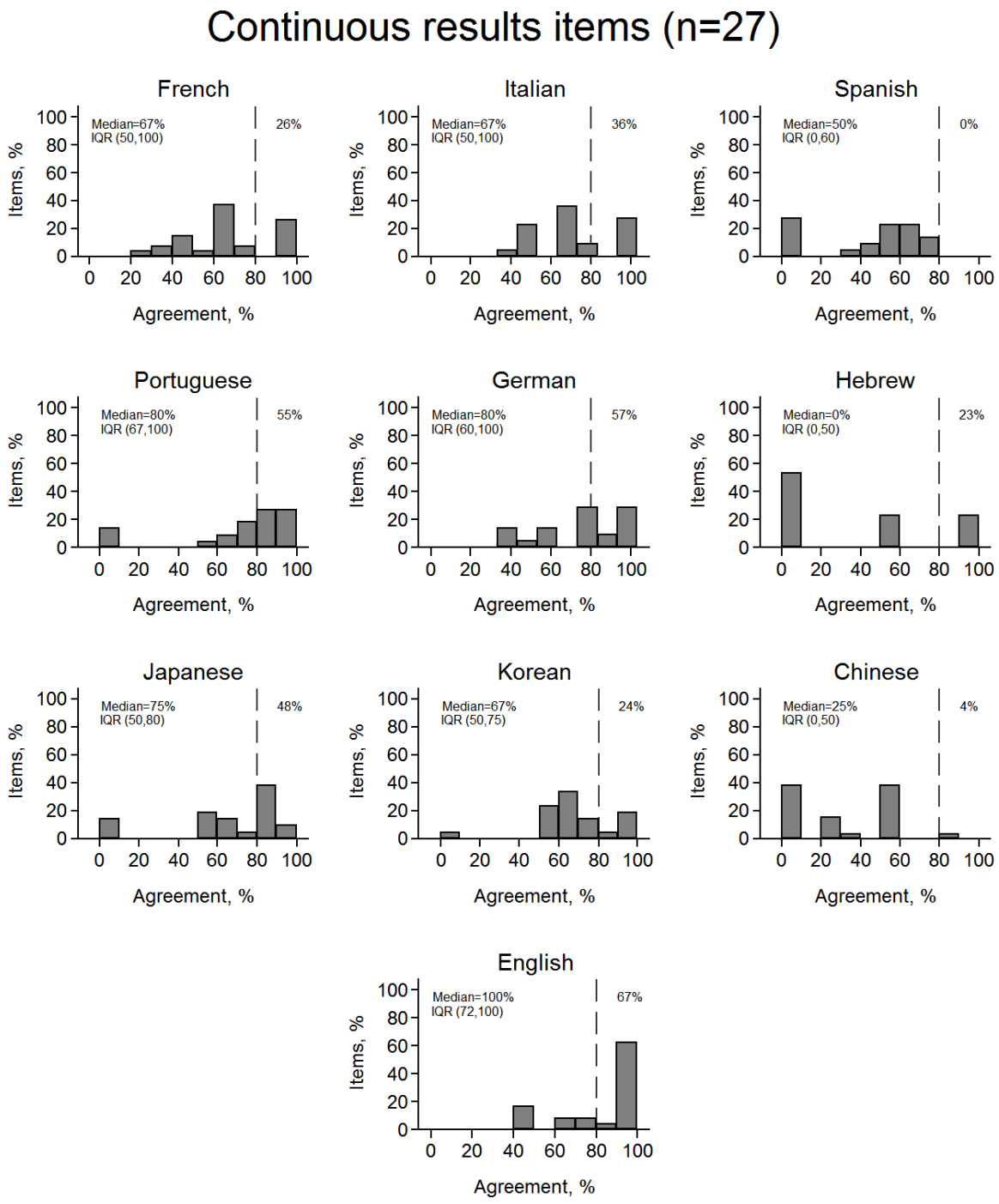

Figure 8. Histograms of the percent agreement for continuous results items, by language. There were a maximum of 27 such items per study. See legend to Figure 1.

Table 4. Percent of items for with there was 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 Results (n=9)		Cont Results (n=27)	
Language	**Median (IQR)***	**P †**	**Median (IQR)***	**P †**	**Median (IQR)***	**P †**	**Median (IQR)***	**P †**	**Median (IQR)***	**P †**	**Median (IQR)***	**P †**	**Median (IQR)***	**P †**
Portuguese	0.85 (0.70, 1.00)	1.00	0.78 (0.71, 0.92)	0.004	0.90 (0.79, 1.00)	0.77	1.00 (0.90, 1.00)	0.17	0.88 (0.71, 1.00)	0.10	0.90 (0.50, 1.00)	0.70	0.80 (0.67, 1.00)	0.091
German	0.80 (0.70, 0.90)	0.68	1.00 (0.82, 1.00)	0.10	1.00 (0.94, 1.00)	0.28	0.90 (0.80, 0.90)	0.58	0.63 (0.57, 0.80)	0.002	0.57 (0.57, 0.57)	0.005	0.80 (0.60, 1.00)	0.084
French	0.94 (0.89, 1.00)	0.44	0.80 (0.69, 0.92)	0.005	0.79 (0.74, 0.85)	0.30	0.80 (0.80, 0.90)	0.34	0.67 (0.67, 1.00)	0.027	0.75 (0.67, 1.00)	0.66	0.67 (0.50, 1.00)	0.007
Korean	0.65 (0.50, 0.80)	0.22	0.83 (0.75, 1.00)	0.006	0.70 (0.65, 0.75)	0.074	0.80 (0.70, 0.90)	0.20	0.71 (0.33, 1.00)	0.028	1.00 (1.00, 1.00)	0.037	0.67 (0.50, 0.75)	0.007
Japanese	0.42 (0.33, 0.50)	0.12	0.71 (0.67, 0.79)	0.001	0.45 (0.37, 0.58)	0.020	0.90 (0.90, 0.90)	0.93	0.63 (0.33, 1.00)	0.019	0.80 (0.50, 1.00)	0.35	0.75 (0.50, 0.80)	0.005
Italian	0.76 (0.67, 0.86)	0.44	0.69 (0.60, 0.89)	<0.001	0.79 (0.74, 0.90)	0.56	0.70 (0.60, 0.90)	0.038	0.89 (0.60, 1.00)	0.040	0.00 (0.00, 1.00)	0.31	0.67 (0.50, 1.00)	0.024
Spanish	0.78 (0.75, 0.80)	0.22	0.83 (0.75, 1.00)	0.025	0.84 (0.71, 0.94)	0.47	0.70 (0.50, 0.70)	0.022	0.80 (0.63, 1.00)	0.034	0.67 (0.50, 0.67)	0.038	0.50 (0.00, 0.60)	<0.001
Hebrew	0.75 (0.75, 0.75)	0.10	0.68 (0.55, 0.78)	0.002	0.71 (0.65, 0.75)	0.080	‡	‡	0.38 (0.00, 1.00)	0.033	0.63 (0.60, 0.90)	0.25	0.00 (0.00, 0.50)	0.001
Chinese	0.42 (0.33, 0.50)	0.12	0.42 (0.27, 0.68)	<0.001	0.56 (0.50, 0.67)	0.041	0.56 (0.44, 0.67)	0.001	0.75 (0.44, 0.89)	0.009	0.57 (0.57, 0.57)	0.003	0.25 (0.00, 0.50)	<0.001
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.

Tx/Cx = treatment and control; Study Char = study characteristics; Dich = dichotomous; Cont = continuous.

Red: P≤0.05; yellow: 0.05<P<0.10; blue: P≥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 5. 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 Results (n=9)		Cont Results (n=27)	
Language	≥80%*, #/N (%)	P †	Median (IQR)*	P †	Median (IQR)*	P †	Median (IQR)*	P †	Median (IQR)*	P †	Median (IQR)*	P †	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%)		11/12 (8%)		3/4 (25%)		8/9 (11%)		7/7 (0%)		6/8 (25%)		16/24 (33%)	

"n" refers to the number of items per section.
Tx/Cx = treatment and control; Study Char = study characteristics; Dich = dichotomous; Cont = continuous.

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.

Discussion

Our results showed that using Google Translate to translate medical articles is a feasible and not resource-intensive process, and leads to operationally workable English versions in many cases. 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 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.

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 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.

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 list of data extraction elements, but could include a larger number of articles per language. A followup study should also perform double (or more) data extraction for both the original language and the translated articles; this would allow a better determination that differences in extraction are due to translation errors, rather than differences among extractors. A formal collaborative study by the EPCs could harness the language skills across the different centers, would enable the 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 than usual to error. 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.

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

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.

2. Committee on Standards for Systematic Reviews of Comparative Effectiveness Research. Finding what works in health care: standards for systematic reviews. Washington, DC. 2011.

3. Heres S, Wagenpfeil S, Hamann J, et al. Language bias in neuroscience--is the Tower of Babel located in Germany? Eur Psychiatry 2004;19(4):230-2.

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.

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.

6. Juni P, Holenstein F, Sterne J, et al. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol 2002;31(1):115-23.

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.

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 ON Canada. 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.

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.

11. Sheppard F. High-resolution computed tomography for clinical imaging of bone microarchitecture. Presse Med 2011;40(6):565-6.