**Discussion** | Discharge instructions were translated by the new GT algorithm with higher accuracy and fewer seriously harmful inaccuracies than previously,[2] yet 2% of Spanish and 8% of Chinese sentence translations had potential for significant harm. While GT can supplement (not replace) written English instructions, machine-translated instructions should include a warning about potentially inaccurate translations.

Clinicians using GT can reduce potential harm by having patients read translations while receiving verbal instructions; being vigilant about spelling and grammar; and avoiding complicated grammar, medical jargon (eg, fingerstick), and colloquial English.

Study limitations include assessment of only 2 languages (though our inclusion of Chinese is a strength, since non-European languages are often less accurately translated by machines); no assessment of translation readability; and no comparison to human translators.

Google Translate can be used to translate clinician-entered, patient-specific ED instructions for Spanish- and Chinese-speaking patients. Potential for harm can be minimized by using clear communication practices. We recommend including English instructions and automated warnings regarding the use of machine translation.

**Elaine C. Khoong, MD, MS**
**Eric Steinbrook, BA**
**Cortlyn Brown, MD**
**Alicia Fernandez, MD**

**Author Affiliations:** Division of General Internal Medicine, Department of Medicine at Zuckerberg San Francisco General Hospital, University of California, San Francisco (Khoong, Fernandez); University of Michigan School of Medicine, Ann Arbor (Steinbrook); Department of Emergency Medicine, University of California, San Francisco (Brown); Center for Vulnerable Populations at University of California, San Francisco (Fernandez).

**Accepted for Publication:** November 13, 2018.

**Corresponding Author:** Elaine C. Khoong, MD, MS, Division of General Internal Medicine, Department of Medicine at Zuckerberg San Francisco General Hospital, University of California, San Francisco, 1001 Potrero Ave, 1M, San Francisco, CA 94122 (elaine.khoong@ucsf.edu).

**Published Online:** February 25, 2019. doi:10.1001/jamainternmed.2018.7653

**Author Contributions:** Dr Khoong had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
*Study concept and design:* Khoong, Brown, Fernandez.
*Acquisition, analysis, or interpretation of data:* All authors.
*Drafting of the manuscript:* All authors.
*Critical revision of the manuscript for important intellectual content:* All authors.
*Statistical analysis:* Khoong, Brown.
*Administrative, technical, or material support:* Steinbrook, Fernandez.
*Study supervision:* Fernandez.

**Conflict of Interest Disclosures:** None reported.

1. Johnson A, Sandford J, Tyndall J. Written and verbal information versus verbal information only for patients being discharged from acute hospital settings to home. *Cochrane Database Syst Rev*. 2003;4(4):CD003716. doi:10.1002/14651858.CD003716

2. Khanna RR, Karliner LS, Eck M, Vittinghoff E, Koenig CJ, Fang MC. Performance of an online translation tool when applied to patient educational material. *J Hosp Med*. 2011;6(9):519-525. doi:10.1002/jhm.898

3. Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. https://arxiv.org/abs/1609.08144. Accessed January 17, 2019.

4. Weiss AJ, Wier LM, Stocks C, Blanchard J. Overview of Emergency Department Visits in the United States, 2011. Agency for Healthcare Research and Quality, Healthcare Cost and Utilization Project Statistical Brief #174. June 2014. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb174-Emergency-Department-Visits-Overview.pdf. Accessed January 17, 2019.

5. Castro CM, Wilson C, Wang F, Schillinger D. Babel babble: physicians' use of unclarified medical jargon with patients. *Am J Health Behav*. 2007;31(suppl 1):S85-S95. doi:10.5993/AJHB.31.s1.11

6. Nápoles AM, Santoyo-Olsson J, Karliner LS, Gregorich SE, Pérez-Stable EJ. Inaccurate language interpretation and its clinical significance in the medical encounters of Spanish-speaking Latinos. *Med Care*. 2015;53(11):940-947. doi:10.1097/MLR.0000000000000422

## Evaluation of the Inclusion of Studies Identified by the FDA as Having Falsified Data in the Results of Meta-analyses: The Example of the Apixaban Trials

Clinical trials under the purview of the US Food and Drug Administration (FDA) have been shown to report falsified data.[1] The FDA warns researchers when falsified data are discovered, but these data have made it into the medical literature.[1,2] The desire to include all available data in a meta-analysis to obtain the "best estimate" of effect size may result in the inclusion of falsified data, which may compromise future research, policy decisions, and patient care.[3] The present study evaluates the inclusion of studies with falsified data in meta-analyses.

**Methods** | As identified by Seife,[2] the ARISTOTLE clinical trial, which studied the pharmaceutical agent apixaban, had the most publications containing falsified data. Therefore, we conducted a systematic review, per the *Cochran Handbook for Systematic Reviews of Interventions* (PROSPERO Identifier: CRD42017055627),[4] to identify meta-analyses that contained at least 1 ARISTOTLE clinical trial publication with falsified data. The institutional review board of Florida International University determined that board approval was not required because the study involved no human participants.

Figure. Study Inclusion Flow Diagram

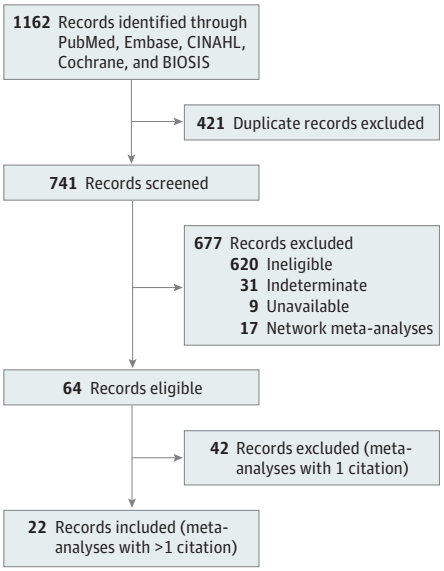

Table. Sensitivity Analysis Estimate Changes

| Meta-analysis Estimate Change | Meta-analyses, No (%) | | |
|---|---|---|---|
| | Subgroup | Full | All |
| OR crossed 1.0 | 3 (15) | 2 (7) | 5 (5) |
| UCI crossed 1.0 | 8 (60) | 14 (43) | 22 (22) |
| OR and UCI crossed 1.0 | 2 (10) | 2 (7) | 4 (4) |
| LCI crossed 1.0 | 1 (5) | NA | 1 (1) |
| Total | 14 (44) | 18 (27) | 32 (32) |

Abbreviations: LCI, 95% lower confidence interval; NA; not applicable; OR, odds ratio; UCI, 95% upper confidence interval.

Sensitivity analyses were conducted to evaluate the inclusion of the ARISTOTLE trial in meta-analyses by determining the odds ratios (ORs) and associated 95% confidence intervals (95% CIs). Estimates were calculated using either a fixed- or random-effects model outlined in each meta-analysis publication using the Mantel-Hanszel method.

**Results** | The ARISTOTLE clinical trial was found in 22 meta-analyses (**Figure**). All meta-analyses found were in English and published between 2012 and 2017. The median number of publications contributing to the analyses was 9 (range, 2-28), and the median meta-analysis publication InCite journal impact factor was 5.658 (range, 3.154-17.202). The median weight of the publication with falsified data toward each meta-analysis was 37.3% (range, 7%-100%).

In our reanalysis of the 22 meta-analyses, we found that 10 (46%) yielded results that would change the initial meta-analysis findings. Each affected meta-analysis had a median of 9.5 publications (range, 2-17), and the median meta-analysis publication InCite journal impact factor was 5.830 (range, 3.154-17.202). The median weight of publications with falsified data was 55.7% (range, 13.1%-99.6%).

From our reanalysis of the 22 meta-analyses, we found that 32 of 99 analyses (32%) yielded results that would change the conclusions of the initial analysis (**Table**). Of the 32 affected estimates, 31 (97%) no longer favored apixaban for the prevention of serious medical issues, and 1 (3%) favored the control.

Of the 99 analyses, 32 (32%) were subgroup analyses, while 67 (68%) were full analyses (Table). Of the 32 subgroup analyses, 14 (44%) yielded results that would change the conclusions of the initial subgroup analyses. For the full analyses, 18 of 67 estimates (27%) had results that were affected.

**Discussion** | This study found that 46% of all meta-analysis publications had conclusions altered by publications with falsified data, and 32% of all the analyses had a considerable change in the outcome. Overall analyses were more robust than subgroup analyses against the effects of publications with falsified data. For ORs not statistically affected, the estimates generally moved toward the null when more than 1 publication remained.

This study was limited to only meta-analyses that contained ARISTOTLE publications identified by Seife to contain falsified data.[2] Not all the data within the ARISTOTLE publications were falsified. However, because the researchers knowingly published falsified data, a form of research misconduct, we removed all ARISTOTLE data.[5]

Our sensitivity analysis results showed that conclusions may be altered in meta-analyses by the inclusion of publications with falsified data. This study should add impetus for robust sensitivity analyses and stronger protections against falsified data. Falsified data can affect not only the original publication, but also any subsequent meta-analyses and any resulting clinical or policy changes resulting from the findings of these studies.

**Craig A. Garmendia, PhD**
**Liliana Nassar Gorra**
**Ana Lucia Rodriguez, MS**
**Mary Jo Trepka, MD, MSPH**
**Emir Veledar, PhD**
**Purnima Madhivanan, MD, PhD**

**Author Affiliations:** Office of Bioresearch Monitoring Operations, Office of Regulatory Affairs, US Food and Drug Administration, Miami, Florida (Garmendia); Department of Epidemiology, Robert Stempel College of Public Health and Social Work, Florida International University, Miami (Nassar Gorra, Trepka, Madhivanan); Department of Psychology, School of Integrated Science and Humanity, Florida International University, Miami (Rodriguez); Department of Biostatistics, Robert Stempel College of Public Health and Social Work, Florida International University, Miami (Veledar).

**Accepted for Publication:** November 8, 2018.

**Corresponding Author:** Craig A. Garmendia, MS, PhD, Department of Epidemiology, Robert Stempel College of Public Health and Social Work, Florida International University, 11200 SW Eighth St, AHC5, 4th Floor, Miami, FL 33199 (cgarm003@fiu.edu).

**Published Online:** March 4, 2019. doi:10.1001/jamainternmed.2018.7661

**Author Contributions:** Dr Garmendia had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
*Study concept and design:* Garmendia, Trepka, Veledar, Madhivanan.
*Acquisition, analysis, or interpretation of data:* Garmendia, Nassar Gorra, Rodriguez.
*Drafting of the manuscript:* Garmendia, Nassar Gorra.
*Critical revision of the manuscript for important intellectual content:* Garmendia, Rodriguez, Trepka, Veledar, Madhivanan.
*Statistical analysis:* Garmendia, Nassar Gorra, Rodriguez, Veledar, Madhivanan.
*Administrative, technical, or material support:* Garmendia.
*Study supervision:* Garmendia, Trepka, Veledar, Madhivanan.

**Conflict of Interest Disclosures:** Dr Garmendia reported grants from Federal Employee Education and Assistance Fund/National Treasury Employee Union during the conduct of the study. Ms Rodriguez reported funding from a predoctoral training program by The Fogarty Institute (Global Health Equity Scholars Program) National Institutes of Health (NIH) FIC D43TW010540. No other disclosures were reported.

**Funding/Support:** Dr Garmendia received funding through the Federal Employee Education and Assistance Fund/National Treasury Employee Union. Ms Rodriguez was funded by The Global Health Equity Scholars Program, National Institutes of Health (NIH) (FIC D43TW010540).

**Role of the Funder/Sponsor:** The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of

the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Disclaimer:** This article reflects the view of the authors and should not be construed to represent the FDA's views or policies.

**Additional Contributions:** We thank Stanislaw F. Wnuk, PhD, of Florida International University, for his collaboration on the design of this study, Barbara M. Sorondo, MS, MLIS, of Florida International University, for her collaboration in the systematic review method, and Katrina Epnere, MPH, of Statistics Collaborative, for her statistical analysis guidance. They received no compensation for their contributions.

1. Garmendia CA, Bhansali N, Madhivanan P. Research misconduct in FDA-regulated clinical trials: a cross-sectional analysis of warning letters and disqualification proceedings. *Ther Innov Regul Sci*. 2018;52(5):592-605.

2. Seife C. Research misconduct identified by the US Food and Drug Administration: out of sight, out of mind, out of the peer-reviewed literature. *JAMA Intern Med*. 2015;175(4):567-577. doi:10.1001/jamainternmed.2014.7774

3. Slavin RE. Best evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol*. 1995;48(1):9-18. doi:10.1016/0895-4356(94)00097-A

4. The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions (Version 5.1.0). https://training.cochrane.org/handbook. Accessed September 5, 2018.

5. Office of Research Integrity, US Department of Health and Human Services. Definition of Research Misconduct (April 25, 2011). https://ori.hhs.gov/definition-misconduct. Accessed September 5, 2018.

## Evidence-Based Medicine and the American Thoracic Society Clinical Practice Guidelines

The American Thoracic Society (ATS) issues clinical practice guidelines for the care of patients with pulmonary and critical care disease. The utility of ATS guidelines depends on the quality of the evidence base underpinning recommendations and whether the guidelines permit the practice of evidence-based medicine (EBM).[1,2] However, the extent to which ATS guidelines are substantiated by high-quality evidence and can be used to promote EBM is unknown.

Editor's Note page 561

**Methods** | Two of 3 investigators (R.C.S., K.D., and A.N.M.) reviewed each ATS clinical practice guideline recommendations listed on the ATS website as of August 1, 2017, that pertained to adults.[3] We abstracted the following domains necessary for evidence-based clinical decision making based on prior conceptual frameworks[1,2]: recommendation type, recommendation strength (using the Grading of Recommendations Assessment, Development, and Evaluation [GRADE] scoring system[4] of strong [benefits clearly outweigh risks in most patients] vs low/conditional [benefits do not clearly outweigh risks in a substantial minority of patients]), quality of evidence (using GRADE categories[4] of high [further research is unlikely to change estimate of effect] to very low [any estimate of effect is uncertain]), EBM measures, and patient context. Institutional review board approval was not needed because no human participants were included.

We defined recommendations as including the basic set of EBM measures if they included at least 1 measure of test performance for diagnostic recommendations (sensitivity, specificity, or likelihood ratio) and at least 1 measure of absolute benefit or harm for therapeutic recommendations (absolute risk reduction/increase, number needed to treat/harm, or relative risk with incidence of the outcome for the control group). For patient context, we ascertained whether the narrative text included any discussion of a person's severity of illness or comorbidities, sociopersonal context, prognosis, or personal preference and how these domains might influence the recommendation.[1] Differences between reviewers were resolved through negotiated consensus, aiming to achieve agreement using the most inclusive definitions. Two-sided $P < .05$ for descriptive statistics indicated significance.

**Results** | Among 222 unique recommendations from 16 separate guidelines, 141 (63.5%) were based on low-quality evidence, whereas fewer than 1 in 10 (19 [8.6%]) were based on high-quality evidence (**Table 1**). Nonetheless, 86 (38.7%) were designated strong recommendations. Higher quality of evidence was associated with an increased probability of receiving a strong recommendation; 29 of 141 low-quality evidence recommendations (20.6%), 41 of 62 moderate-quality evidence recommendations (66.1%), and 16 of 19 high-quality evidence recommendations (84.2%) were strongly recommended ($P < .001$ for trend). However, most strong recommendations were not supported by high-quality evidence (16 of 86 [18.6%]).

Of 52 diagnostic testing recommendations, 26 (50.0%) presented the test's sensitivity, specificity, or likelihood ratios. Of 165 therapeutic recommendations, 76 (46.1%) reported the

**Table 1. Summary of the Evidence Base for ATS Clinical Practice Guidelines**

| Recommendation Type | Recommendations, No. | Recommendations, No. (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Strength[a] | | Quality of Evidence[b] | | | Meets Basic Definition | |
| | | Strong | Low/Conditional | High | Medium | Low | EBM Measures | Patient Context |
| Overall[c] | 222 | 86 (38.7) | 136 (61.3) | 19 (8.6) | 62 (27.9) | 141 (63.5) | 102 (45.9) | 101 (45.5) |
| Diagnostic | 52 | 19 (36.5) | 33 (63.5) | 1 (1.9) | 16 (30.8) | 35 (67.3) | 26 (50.0) | 3 (5.8) |
| Therapeutic | 165 | 65 (39.4) | 100 (60.6) | 18 (10.9) | 44 (26.7) | 103 (62.4) | 76 (46.1) | 98 (59.4) |

Abbreviations: ATS, American Thoracic Society; EBM, evidence-based medicine.

[a] Classified using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) score except the guideline for community-acquired pneumonia (CAP), which used strong (most patients should receive the intervention), medium, and weak (many health care professionals would not follow this recommendation) categories. CAP recommendations of medium or weak were reclassified as low/conditional.

[b] Classified using the 4 GRADE categories, combining low- and very-low-quality ratings into a single low category, given the similar uncertainty and because several ATS guidelines only used 1 of these categories.

[c] Includes diagnostic, therapeutic, screening (n = 3), and monitoring (n = 3) recommendations. One recommendation included both a therapeutic and diagnostic recommendation.