

ORIGINAL ARTICLE

Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses

Richeek Pradhan^a, David C. Hoaglin^a, Matthew Cornell^a, Weisong Liu^b, Victoria Wang^c, Hong Yu^{b,c,d,e,*}

^aDepartment of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

^bDepartment of Computer Science, University of Massachusetts Lowell, Lowell, MA, USA

^cDepartment of Medicine, UMass Medical School, Worcester, MA, USA

^dDepartment of Computer Science, University of Massachusetts Amherst, Amherst, MA, USA

^eCenter for Healthcare Organization & Implementation Research, Edith Nourse Rogers Memorial Veterans Hospital, Bedford, MA, USA

Accepted 19 August 2018; Published online 23 September 2018

Abstract

Objectives: Systematic reviews and meta-analyses are labor-intensive and time-consuming. Automated extraction of quantitative data from primary studies can accelerate this process. ClinicalTrials.gov, launched in 2000, is the world's largest trial repository of results data from clinical trials; it has been used as a source instead of journal articles. We have developed a Web application called EXACT (EXtracting Accurate efficacy and safety information from ClinicalTrials.gov) that allows users without advanced programming skills to automatically extract data from ClinicalTrials.gov in analysis-ready format. We have also used the automatically extracted data to examine the reproducibility of meta-analyses in three published systematic reviews.

Study Design and Setting: We developed a Python-based software application (EXACT) that automatically extracts data required for meta-analysis from the ClinicalTrials.gov database in a spreadsheet format. We confirmed the accuracy of the extracted data and then used those data to repeat meta-analyses in three published systematic reviews. To ensure that we used the same statistical methods and outcomes as the published systematic reviews, we repeated the meta-analyses using data manually extracted from the relevant journal articles. For the outcomes whose results we were able to reproduce using those journal article data, we examined the usability of ClinicalTrials.gov data.

Results: EXACT extracted data at ClinicalTrials.gov with 100% accuracy, and it required 60% less time than the usual practice of manually extracting data from journal articles. We found that 87% of the data elements extracted using EXACT matched those extracted manually from the journal articles. We were able to reproduce 24 of 28 outcomes using the journal article data. Of these 24 outcomes, we were able to reproduce 83.3% of the published estimates using data at ClinicalTrials.gov.

Conclusion: EXACT (<http://bio-nlp.org/EXACT>) automatically and accurately extracted data elements from ClinicalTrials.gov and thus reduced time in data extraction. The ClinicalTrials.gov data reproduced most meta-analysis results in our study, but this conclusion needs further validation. © 2018 Elsevier Inc. All rights reserved.

Keywords: ClinicalTrials.gov; Meta-analysis; Reproducibility; Automatic data extraction; Systematic review; Simeprevir; Trametinib; Vortioxetine

1. Introduction

Systematic reviews and meta-analyses can provide the highest level of medical evidence, influencing clinical guidelines and regulatory decision-making [1]. Conducting systematic reviews requires domain experts to manually search and evaluate primary studies, decide which outcomes are to be evaluated, extract the data, and statistically

synthesize the data, making the entire process labor-intensive, time-consuming, and expensive [2]. For example, the median time from publication of protocols of Cochrane systematic reviews to final publication of the systematic review is 2.4 years [3]. Often, the pace of research renders a systematic review out-of-date by the time it is published [4]. However, disease management guidelines, published every few years, depend on up-to-date systematic reviews. Moreover, therapeutic domains that require expeditious decision-making (e.g., a developing public health emergency) may not accommodate a protracted evidence-synthesis process [5]. Further, manual data extraction is

Conflict of interest statement: None.

* Corresponding author. Tel.: 978-934-3620; fax: 978-934-3551.

E-mail address: hong_yu@uml.edu (H. Yu).

What is new?**Key findings**

- We developed a web-based application called EXACT that automatically and accurately extracts data from ClinicalTrials.gov. We reproduced results of published meta-analyses with data extracted using EXACT.

What this adds to what was known?

- Though searches of registries are recommended as a part of systematic reviews to identify unpublished data, no consensus exists on usability of trial data reported at registries. We show that from using trial results reported at ClinicalTrials.gov registry one can reproduce most results of published meta-analyses. Using a tool like EXACT to extract such data saves time.

What is the implication and what should change now?

- Meta-analyses could use data from ClinicalTrials.gov extracted using EXACT instead of primary article data.

inefficient and error-prone [6,7]; usually, two reviewers independently extract data from each included study, but errors may still remain. Thus, there is a growing consensus on the need to make the systematic review faster, less error-prone, and more efficient [5].

One strategy to accelerate evidence synthesis in systematic reviews automates various steps of the process by applying computation technology. Current methods focus on natural language processing approaches that automatically extract information from abstracts and full texts of published primary studies [2,8]. Most of these algorithms, however, extract only baseline data elements such as the demographic characteristics of the patients, the clinical indication, the nature of the interventions, the study design, and the names of the outcomes studied. Other systems automatically assess “risk of bias” in primary studies by summarizing study design characteristics [9]. Although extraction of baseline data would partly reduce data extraction time, mistakes in extracting these elements are less common [7]. Fewer attempts have been made to extract the *number* of patients having specific outcomes. This is primarily because of varied journal formats and lack of uniformity in reporting language. For example, studies using machine learning to extract numbers of patients with outcomes, from just the article abstracts (thus on a very small number of outcomes), reported poor performance (positive predictive value about

69%) [10,11]. Thus, current automatic methods seem to introduce extraction errors of their own.

To bypass the error-prone machine learning techniques while still automating quantitative data extraction from primary studies, one could tap alternative sources of trial data such as the registry ClinicalTrials.gov. Established as a result of the Food and Drug Administration (FDA) Modernization Act in 1997 and launched in 2000, ClinicalTrials.gov is the largest trial registry in the world [12]. Its standard reporting format makes the registry database more amenable to automatic extraction than journals, which follow various reporting conventions. Previous initiatives for ClinicalTrials.gov data include Sherlock [13], a proprietary system to extract ClinicalTrials.gov trials focusing on pain outcomes into an analysis-ready database, and the Clinical Trials Transformation Initiative (CTTI) [14], a public–private initiative that has organized the available information in the ClinicalTrials.gov database, but extracting trial data from the CTTI database requires advanced programming skills. Thus, a user-friendly and publicly available method for extracting extensive trial data available in ClinicalTrials.gov was not available.

To fill this gap, we describe the development of EXACT (EXtracting Accurate efficacy and safety information from ClinicalTrials.gov: <http://bio-nlp.org/EXACT/>), an automatic data-extraction application that produces analysis-ready spreadsheets consisting of data from trial results available at ClinicalTrials.gov. We assess its accuracy by comparing the data extracted by EXACT with the data at ClinicalTrials.gov. Finally, we use the extracted data to repeat meta-analyses from three published systematic reviews. In the process, we explore whether the use of ClinicalTrials.gov data substantially alters conclusions of meta-analyses.

2. Methods

2.1. Development of EXACT

We developed the Web-based application EXACT in two parts: (1) a Python-based library to parse records in the ClinicalTrials.gov database (the library contains 30 functions in categories corresponding to baseline information [trial title, study type, conditions, interventions, and design] and routines to extract data from the sections on baseline, outcome, participant flow, and reported events) and (2) an application that allows the user to specify the data desired. Programming details on the development of EXACT are available in Appendix 1. We used the Extensible Markup Language (XML) files available from the search list as the source of the data for the EXACT library [15].

A user of EXACT can initiate a search with the unique ClinicalTrials.gov identifier for a particular trial. Then, if the trial has reported results at ClinicalTrials.gov, the user can download all trial data or specify the quantitative data

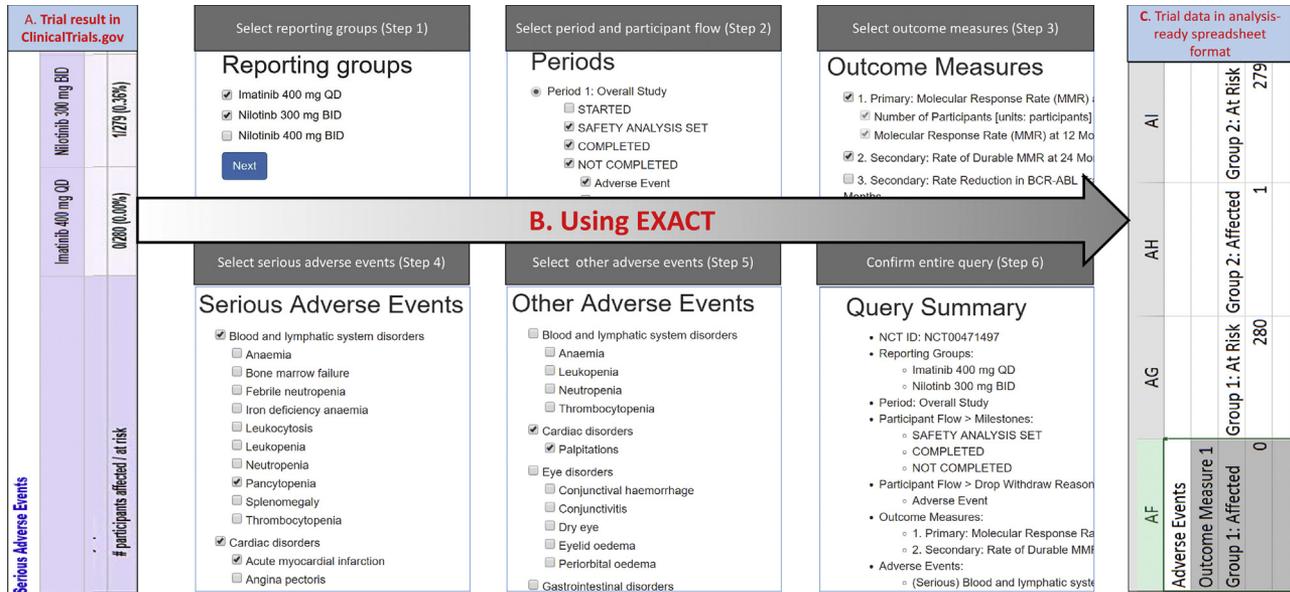


Fig. 1. EXACT's user interface. (A) Screenshot of trial result reported at ClinicalTrials.gov. (B) Screenshot of six steps through which the user specifies the items to be downloaded (Appendix 2 provides a user manual). (C) Screenshot of data extracted in Excel format.

to be extracted by selecting (1) the reporting groups, (2) period (main trial period/follow-up period) and participant flow, (3) outcome measures, (4) serious adverse events, and (5) other adverse events. In addition, qualitative trial data including trial title, study type, conditions, interventions, and design are always downloaded in the same spreadsheet with no need to select them. Figure 1 shows the user interface, and Appendix 2 is a user manual for EXACT.

2.2. Initial validation of data extracted by EXACT

In the development phase, to ensure proper functioning and fidelity of the extracted data, we validated all library functions via 25 unit tests (testing the individual components and functions of the software). Test inputs were eight actual ClinicalTrials.gov XML files that represented a range of trials, along with hand-collected expected outputs for them, obtained from each trial's ClinicalTrials.gov results page.

2.3. Validation of use of EXACT in meta-analysis research

To validate use of EXACT in conducting meta-analyses, we selected published systematic reviews from the literature, listed the outcomes in their meta-analyses, used EXACT to extract the corresponding data from ClinicalTrials.gov, and compared those data with the data reported on the ClinicalTrials.gov Website. We also manually extracted the outcome data from the journal articles for the trials used in the systematic reviews. First, we attempted to reproduce the meta-analyses in the published systematic reviews using the manually extracted data, thus ensuring

that we were using the same methods as the systematic review authors. Then we used these validated methods and ClinicalTrials.gov data extracted by EXACT to reproduce the meta-analyses in the published systematic reviews. Finally, we compared the meta-analysis results of the published systematic reviews and the results of repeating the meta-analyses using the data extracted by EXACT. If data at ClinicalTrials.gov are to be used in meta-analyses, then we should be able to reproduce most meta-analysis results obtained by usual practice.

2.3.1. Selection of published systematic review articles

Because the requirement to report trial results at ClinicalTrials.gov is relatively recent [16], we searched for systematic reviews of drugs recently approved by FDA [17]. From the 27 new molecular entities (NMEs) approved by FDA in 2013, we used a random number generator to select a random sample of three NMEs (approximating 10%). We then searched PubMed for systematic reviews involving those drugs. For this proof-of-concept investigation, we selected one meta-analysis article per drug. We excluded systematic reviews that (1) used Bayesian or other model-based methods, (2) did not include efficacy or safety outcomes, (3) did not have trial results available in published articles and at ClinicalTrials.gov, or (4) had ambiguous outcomes. When more than one article for a drug satisfied all these criteria, we selected the one that had the largest number of outcomes.

2.3.2. Data extraction

We listed the outcomes examined in the three selected systematic reviews and extracted the relevant quantitative data in two ways: (1) by EXACT from the corresponding

ClinicalTrials.gov entries and (2) manually from the journal articles used in the systematic reviews. We recorded the time required for each method and checked for any mismatch between the downloaded data and the source.

2.3.3. Internal validation of the accuracy of data extracted by EXACT

To evaluate the accuracy of the application, we manually matched the data extracted by EXACT with the data at ClinicalTrials.gov. Although we had already validated the accuracy of EXACT in the development phase (Section 2.2) by unit tests, the present step confirmed the extraction for 15 trials involving three randomly selected drugs. This was important as we had little control over which trials would come up for data extraction.

2.3.4. Confirming the meta-analysis methods with the manually extracted data

To ensure that we used the same methods as the published systematic reviews, we first repeated the meta-analyses of outcomes in the systematic reviews using the data manually extracted from the primary study articles. We considered a published systematic review to have been reproduced if (1) our overall point estimate of effect size (relative risk/odds ratio/hazard ratio) was within $\pm 20\%$ of the point estimate in the published systematic review and (2) the *P*-value remained on the same side of 0.05 as in the published systematic review. We adhered to the systematic review's measure of effect (relative risk for 25 outcomes, odds ratio for one outcome, and hazard ratio for

two outcomes). The 20% margin allowed for differences in numerators and denominators in both test and control arms. All analyses used Stata version 14 (StataCorp LP, College Station, TX, USA) and used the same random-effects or fixed-effect method as in the published systematic review (illustrative Stata codes in [Appendix 3](#)).

2.3.5. Repeating the meta-analyses with the data extracted by EXACT

After confirming that we were using the same methods as the published systematic reviews, we repeated the meta-analyses using the data extracted by EXACT. We considered a meta-analysis to be reproduced if (1) the point estimate of effect size was within $\pm 20\%$ of the point estimate in the published systematic review and (2) the *P*-value remained on the same side of 0.05 as in the published systematic review.

3. Results

From the 27 NMEs approved by the FDA in 2013, we randomly selected three: simeprevir, trametinib, and vortioxetine. [Figure 2](#) describes the article selection, [Table 1](#) describes the articles selected (Qu et al., 2016 [18], six trials; Abdel-Rahman et al., 2016 [19], four trials; and Li et al., 2016 [20], five trials), and [Appendix 4](#) describes the reasons for excluding other systematic review articles.

The three systematic review articles contained meta-analyses of a total of 28 outcomes. We spent approximately

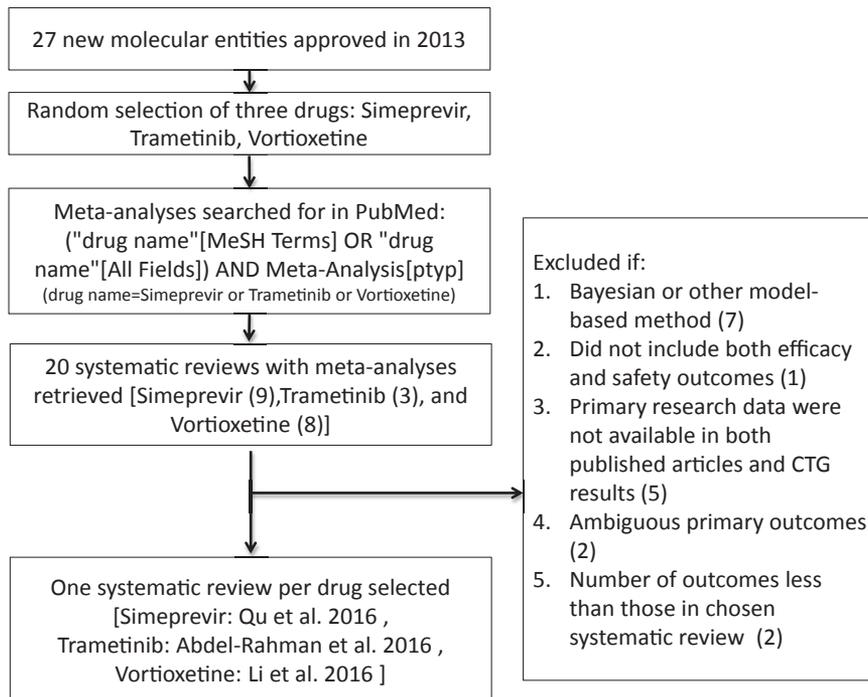


Fig. 2. Flowchart for selection of meta-analyses repeated using data manually extracted from original articles and data extracted by EXACT from ClinicalTrials.gov.

Table 1. Selected systematic reviews

Systematic review	Comparison	Number of RCTs pooled	Outcomes selected for meta-analysis
Qu et al., 2016	Comparison of simpeprevir + peginterferon + ribavirin versus peginterferon + ribavirin in HCV infection	6	Sustained virological response at 12 weeks, rapid virological response, serious adverse events, discontinuation
Abdel-Rahman et al., 2016	Comparison of BRAF inhibitors + trametinib versus BRAF inhibitors in BRAF-mutated melanoma	4	Hazard ratio for progression-free survival, hazard ratio for overall survival, overall response rate, ^a diarrhea, hypertension, decreased ejection fraction, acneiform dermatitis, pyrexia, squamous cell carcinoma
Li et al., 2016	Vortioxetine versus duloxetine in major depressive disorder	5	Response rate, remission rate, ^a nausea, constipation, Hyperhidrosis, diarrhea, dizziness, dry mouth, fatigue, insomnia, somnolence, nasopharyngitis, ^a decreased appetite, ^a headache, vomiting

Abbreviations: BRAF, a gene encoding B-Raf protein; HCV, hepatitis C virus; RCT, randomized controlled trial.

^a Outcomes in published systematic reviews that could not be reproduced using article data by us.

10 hours extracting data from the 15 journal articles. Although EXACT automatically extracts the data elements, we manually convert certain formats (e.g., from percentages to numbers of patients with events), and therefore, we spent approximately 4 hours in using EXACT. Thus, overall, EXACT reduced the time of data extraction by 60%. On manual comparison, 100% of the data elements extracted using EXACT matched the data posted at ClinicalTrials.gov.

Using the data manually extracted from the journal articles, we were able to reproduce results for 24 outcomes. (Details of the other four outcomes and the possible reasons for being unable to reproduce them are in [Appendix 5](#). In one instance, it seems clear that we were unable to reproduce a result because the authors did not actually use the stated method.) The median point estimate reported in the published systematic review for the five efficacy outcomes was 0.83 (range 0.56–9.57), and that for the 19 safety outcomes was 0.70 (range 0.16–4.63).

Having ensured that we were using the same statistical methods as the published systematic reviews, we sought to reproduce the meta-analysis results for these 24 outcomes using the data extracted by EXACT. The 24 outcomes required extraction of 482 pairs of data elements from original articles and ClinicalTrials.gov via EXACT. The percentage of extracted numbers matched between the two sources was 86.8% (details in [Appendix 6](#)). An equal amount of trial data was available from the original articles and ClinicalTrials.gov for 19 of the 24 outcomes. More trial data were available from original articles for one efficacy outcome (hazard ratio for overall survival in trametinib), and ClinicalTrials.gov provided more data for four safety outcomes (hypertension and acneiform dermatitis for trametinib, hyperhidrosis and somnolence for vortioxetine).

Using the data extracted from ClinicalTrials.gov, we were able to reproduce results for 20 of the 24 outcomes

(83.3%). (Among all 28 outcomes, 21 [75%] were reproduced.) The results are shown in [Table 2](#). We were unable to reproduce four safety outcomes (discontinuation in Qu et al. 2016: deviation from published systematic review –22.2%; hypertension in Abdel-Rahman et al. 2016: deviation 8.2%, *P*-value change from 0.07 in published systematic review to 0.005 in reproduction; decreased ejection fraction in Abdel-Rahman et al. 2016: deviation –27.6% from published systematic review; squamous cell carcinoma in Abdel-Rahman et al. 2016: deviation 25.0% from published systematic review, *P*-value change from < 0.00001 in published systematic review to < 0.001 in reproduction). For the five efficacy outcomes, the percentage differences in point estimates between reproduction and published systematic reviews were –4.0%, –3.6%, 0%, +0.1%, and +1.1%; and for the 15 safety outcomes, the median was –0.76% (interquartile range –2.46% to 0%, minimum –17.14%, maximum 10.10%).

(We were unable to reproduce meta-analyses for weighted mean differences in Li et al. 2016 because the journal articles for this systematic review did not consistently report measures of dispersion, [Appendix 7](#). The meta-analyses using data at ClinicalTrials.gov, where standard deviations were reported, are in [Appendix 8](#). [Appendix 9](#) describes the data extracted manually and using EXACT.)

4. Discussion

We developed the Web-based application EXACT to automatically extract data from ClinicalTrials.gov. This application accurately extracts both baseline data and quantitative outcome data and requires less time than current usual practice (i.e., manual extraction from journal articles). We validated the use of ClinicalTrials.gov data extracted using EXACT by repeating the meta-analyses in

Table 2. Results of meta-analyses from published systematic review and meta-analyses using data manually extracted from original articles and from ClinicalTrials.gov using EXACT

Outcomes	Type of outcome	Point estimates of effect size (relative risk/hazard ratio ^a) and 95% confidence limits			P values		
		Published systematic review	Data from original articles	Data from CTG via EXACT	Published systematic review	Data from original articles	Data from CTG via EXACT
Qu et al.2016, simeprevir							
Sustained virological response at 12 weeks	Efficacy	1.69 (1.37–2.08)	1.69 (1.37–2.08)	1.69 (1.37–2.08)	<0.001	<0.001	<0.001
Rapid virological response	Efficacy	9.57 (5.82–15.73)	9.57 (5.82–15.73)	9.68 (5.88–15.95)	<0.001	<0.001	<0.001
Serious adverse events	Safety	0.67 (0.47–0.94)	0.67 (0.47–0.94)	0.65 (0.46–0.92)	0.023	0.024	0.017
Discontinuation*	Safety	1.26 (0.58–2.74)	1.033 (0.65–1.73)	0.98 (0.44–2.19)	0.566	0.899	0.970
Abdel-Rahman et al., 2016, Trametinib							
Hazard ratio for progression-free survival ^a	Efficacy	0.56 (0.49–0.64)	0.54 (0.47–0.62)	0.54 (0.47–0.62)	<0.00001	<0.001	<0.001
Hazard ratio for overall survival ^a	Efficacy	0.7 (0.58–0.84)	0.57 (0.46–0.68)	0.67 (0.52–0.83)	0.00001	<0.001	<0.001
Diarrhea	Safety	1.3 (1.3–1.49)	1.3 (1.13–1.48)	1.29 (1.13–1.47)	0.0002	<0.001	<0.001
Hypertension*	Safety	1.22 (0.99–1.52)	1.22 (0.98–1.51)	1.32 (1.08–1.62)	0.07	0.068	0.005
Decreased ejection fraction*	Safety	4.63 (2.56–8.37)	4.63 (2.56–8.36)	3.35 (2.02–5.55)	<0.00001	<0.001	<0.001
Acneiform dermatitis	Safety	1.61 (1.03–2.53)	1.61 (1.02–2.53)	1.58 (1.12–2.21)	0.04	0.038	0.008
Pyrexia	Safety	1.98 (1.72–2.27)	1.97 (1.71–2.27)	2.18 (1.91–2.49)	<0.00001	<0.001	<0.001
Squamous cell carcinoma*	Safety	0.16 (0.1–0.25)	0.16 (0.1–0.25)	0.2 (0.11–0.36)	<0.00001	<0.001	<0.001
Li et al., 2016, Vortioxetine							
Response rate	Efficacy	0.83 (0.77–0.89)	0.83 (0.78–0.89)	0.83 (0.77–0.89)	<0.001	<0.001	<0.001
Nausea	Safety	0.7 (0.56–0.87)	0.75 (0.61–0.94)	0.7 (0.61–0.81)	0.001	0.01	<0.001
Constipation	Safety	0.47 (0.34–0.64)	0.45 (0.32–0.65)	0.45 (0.32–0.65)	<0.001	<0.001	<0.001
Hyperhidrosis	Safety	0.35 (0.23–0.55)	0.31 (0.19–0.5)	0.29 (0.18–0.45)	<0.001	<0.001	<0.001
Diarrhea	Safety	0.74 (0.57–0.97)	0.72 (0.53–0.97)	0.72 (0.53–0.97)	0.03	0.03	0.03
Dizziness	Safety	0.51 (0.37–0.69)	0.51 (0.33–0.8)	0.51 (0.33–0.8)	<0.001	0.004	0.004
Dry mouth	Safety	0.5 (0.39–0.63)	0.48 (0.35–0.65)	0.52 (0.31–0.85)	<0.001	<0.001	0.01
Fatigue	Safety	0.45 (0.32–0.64)	0.44 (0.29–0.67)	0.44 (0.29–0.67)	<0.001	<0.001	<0.001
Insomnia	Safety	0.65 (0.46–0.92)	0.64 (0.42–0.96)	0.64 (0.42–0.96)	0.016	0.033	0.033
Somnolence	Safety	0.33 (0.21–0.52)	0.31 (0.19–0.5)	0.33 (0.21–0.5)	<0.001	<0.001	<0.001
Headache	Safety	0.93 (0.77–1.13)	0.93 (0.74–1.16)	0.93 (0.74–1.16)	0.468	0.523	0.523
Vomiting	Safety	0.7 (0.45–1.09)	0.72 (0.45–1.16)	0.72 (0.45–1.16)	0.110	0.176	0.176

Abbreviations: NA, not available.

Outcomes marked with * could not be reproduced using ClinicalTrials.gov data.

^a Hazard ratios were hand-extracted from ClinicalTrials.gov because EXACT does not support extraction of hazard ratios (these data elements are reported in the statistics section of the ClinicalTrials.gov trial reports whereas EXACT currently extracts quantitative data from the Results sections).

published systematic reviews. Although the data published in the original articles and the data at ClinicalTrials.gov did not always match, the differences seldom qualitatively altered the results of meta-analyses.

One of the most protracted and error-prone steps in conducting systematic reviews involves extraction of data from reports on primary trials [7,21,22]. Inaccuracies in extracting numerical data pose a challenge to the validity and

reproducibility of results [23], particularly amid concerns about “the mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses” [24]. Although automating data extraction from journal articles could possibly expedite the systematic review process, current state-of-the-art machine-learning techniques introduce their own extraction errors, which may deter systematic review researchers from using them. On the other hand, ClinicalTrials.gov, already a recognized alternative source of trial data, presents an easier solution to automation because of its uniform reporting format. As validated in the present study, EXACT fills the gap, making data extraction from ClinicalTrials.gov substantially faster. Moreover, since data extraction by EXACT relies on data mining rather than machine learning techniques, EXACT makes no extraction errors. Because of the error-free extraction and because the reporting format in ClinicalTrials.gov is fairly uncomplicated (compared with data in published articles, which often appear only in text), EXACT could obviate the need for two expert reviewers to extract numerical data, once the outcomes to be examined in the systematic review have been selected [22]. EXACT could also be used by journal reviewers and regulatory authorities to check the ClinicalTrials.gov site for the validity of data entered [25] and match ClinicalTrials.gov reports with the published articles and regulatory submissions corresponding to these studies. This capability will add to research transparency and reproducibility, areas of specific concern at present [26]. Finally, EXACT can serve as a model for developing applications to extract data from other clinical trial registries that harbor results databases (like the European Union Clinical Trials Register).

Using data at ClinicalTrials.gov has many advantages. First, it is an important resource to search for trial data, as only half of the trials posted in the registry are published within the next 4 years [27]. Second, besides including all primary outcomes, ClinicalTrials.gov has been shown to include a wider range of safety data than journal articles [28]. Moreover, by asking trialists to follow the MedDRA terminology to report adverse events, ClinicalTrials.gov imposes definitional uniformity in safety reporting. Third, it is freely available, an important advantage in times of rising journal subscription costs [29].

However, using ClinicalTrials.gov as the *only* data source has several disadvantages. First, the data at ClinicalTrials.gov have minor differences from the data in journal articles, as confirmed in the present study [30,31]. However, recent research shows that, if FDA medical reviews are considered the reference standard for data accuracy, ClinicalTrials.gov and journal articles differ from the FDA reviews at a similar rate, making data at ClinicalTrials.gov no less trustworthy than the journal articles (particularly because the peer review process in journals, where reviewers do not have access to individual patient data, cannot guarantee data accuracy) [32].

Second, reporting of study results at ClinicalTrials.gov is still low [16]. Importantly, however, Fain et al., 2017 have shown that the number of trials reported exclusively at ClinicalTrials.gov is higher than the number of trials reported exclusively in journal articles, making ClinicalTrials.gov a critical source for trial data [33]. Moreover, with the recently promulgated “Final Rule” on ClinicalTrials.gov reporting, legally mandated reporting at ClinicalTrials.gov will encompass a wider variety of trials than before [34].

Third, ClinicalTrials.gov, although it contains a segment on the study design in trial reports, may not contain all information extracted from qualitative evaluation of primary studies, a more nuanced assessment of which may still be available only from journal articles.

Fourth, ClinicalTrials.gov focuses on clinical trials conducted in the United States, considerably narrowing the population of studies that will ever be available for data extraction. For example, 15 World Health Organization—recognized trial registries (apart from ClinicalTrials.gov) cover various georegulatory areas in the world. These registries have between 0% and 33% of trials in common with ClinicalTrials.gov, highlighting that a substantial percentage of trials worldwide remain uncovered if ClinicalTrials.gov is used as the only data source [35].

Thus, at present, searching the literature must accompany the use of an application like EXACT. A workable balance could identify primary studies using conventional bibliographic and registry search, perform risk-of-bias assessment by taking into account information from both journal articles and registry reports, and finally automatically extract quantitative data for meta-analyses. Importantly, this work, by demonstrating that study registries can become the source of data for systematic review and meta-analysis research and also expedite it by automation, calls for more-accurate documentation of data and increased registration and reporting rates at ClinicalTrials.gov and other registries to increase their usability for secondary research.

4.1. Limitations

Our work has several limitations. To validate the use of EXACT in extracting data from ClinicalTrials.gov, we used just three systematic reviews with a total of 15 randomized controlled trials and 28 outcomes. Although the drugs were randomly selected, the sample of the systematic reviews is small; thus, not finding differences between meta-analysis outcomes might merely be because of insufficient sample size. Moreover, for comparison, we only included those systematic reviews where the primary studies were reported in both journal articles and ClinicalTrials.gov. This scenario may not be generalizable because of the low reporting rates, both at ClinicalTrials.gov and in journal articles. Also, we only examined systematic reviews of newly approved drugs, where data receive greater scrutiny, resulting in more-consistent reporting of results. Hence, the similarity of meta-analysis results between ClinicalTrials.gov and

journal articles may not generalize to all drugs. Importantly, although all systematic reviews specifically stated that they looked at journal articles for primary data, we cannot exclude the possibility that the authors obtained some data from ClinicalTrials.gov. On the technical side, use of EXACT software may require training time that, on initial use, partially offsets the reduction in time for data extraction. Finally, future work will need to validate the usefulness of data at ClinicalTrials.gov, in general, and of EXACT, in particular, by using it prospectively on a larger sample of systematic reviews.

5. Conclusion

EXACT helps meta-analysis research by automatically extracting results data from ClinicalTrials.gov. Data from ClinicalTrials.gov reproduced results published in the original systematic reviews in our proof-of-concept analysis.

Acknowledgments

Funding: This work was supported in part by grant HL125089 from the National Institutes of Health (NIH). The contents of this article do not represent the views of NIH.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2018.08.023>.

References

- [1] Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997;126:376–80.
- [2] Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev* 2015;4:78.
- [3] Tricco AC, Brehaut J, Chen MH, Moher D. Following 411 Cochrane protocols to completion: a retrospective cohort study. *PLoS One* 2008;3:e3684.
- [4] Cochrane handbook for systematic reviews of interventions [internet]. Available at <http://handbook.cochrane.org/>. Accessed June 17, 2016.
- [5] Tsertsvadze A, Chen Y-F, Moher D, Sutcliffe P, McCarthy N. How to conduct systematic reviews more expeditiously? *Syst Rev* 2015;4:160.
- [6] Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009;62:1013–20.
- [7] Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* 2010;63:289–98.
- [8] Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev* 2014;3:74.
- [9] Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016;23:193–201.
- [10] Zhao J, Bysani P, Kan M-Y. Exploiting classification correlations for the extraction of evidence-based practice information. *AMIA Annu Symp Proc* 2012;2012:1070–8.
- [11] Summerscales R, Argamon S, Bai S, Hupert J, Shwartz A. Automatic summarization of results from clinical trials. *Bioinformatics and Biomedicine (BIBM)*. IEEE International Conference on 2011, Atlanta, GA, USA, 2011:372–7.
- [12] Home - ClinicalTrials.gov [internet]. Available at <https://clinicaltrials.gov/>. Accessed June 17, 2016.
- [13] Cepeda MS, Lobanov V, Berlin JA. From ClinicalTrials.gov trial registry to an analysis-ready database of clinical trial results. *Clin Trials* 2013;10(2):347–8.
- [14] Home | clinical trials transformation initiative [internet]. Available at <http://www.ctti-clinicaltrials.org/>. Accessed June 17, 2016.
- [15] Downloading content for analysis - ClinicalTrials.gov [internet]. Available at <https://clinicaltrials.gov/ct2/resources/download>. Accessed June 17, 2016.
- [16] Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials.gov. *N Engl J Med* 2015;372:1031–9.
- [17] FDA. Available at <http://wayback.archive-it.org/7993/20161022052135/http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugInnovation/ucm381263.htm>. Accessed October 19, 2017.
- [18] Qu Y, Li T, Wang L, Liu F, Ye Q. Efficacy and safety of simeprevir for chronic hepatitis virus C genotype 1 infection: a meta-analysis. *Clin Res Hepatol Gastroenterol* 2016;40(2):203–12.
- [19] Abdel-Rahman O, ElHalawani H, Ahmed H. Doublet BRAF/MEK inhibition versus single-agent BRAF inhibition in the management of BRAF-mutant advanced melanoma, biological rationale and meta-analysis of published data. *Clin Transl Oncol* 2016;18(8):848–58.
- [20] Li G, Wang X, Ma D. Vortioxetine versus duloxetine in the treatment of patients with major depressive disorder: a meta-analysis of randomized controlled trials. *Clin Drug Investig* 2016;36(7):509–17.
- [21] Jones AP, Remington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. *J Clin Epidemiol* 2005;58:741–2.
- [22] Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. *J Clin Epidemiol* 2006;59:697–703.
- [23] Götzsche PC, Hróbjartsson A, Maric K, Tendam B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* 2007;298:430–7.
- [24] Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q* 2016;94(3):485–514.
- [25] Mayo-Wilson E, Li T, Fusco N, Dickersin K, MUDS Investigators. Practical guidance for using multiple data sources in systematic reviews and meta-analyses (with examples from the MUDS study). *Res Synth Methods* 2018;9(1):2–12.
- [26] Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med* 2016;8(341):341ps12.
- [27] Riveros C, Dechartres A, Perrodeau E, Haneef R, Boutron I, Ravaut P. Timing and completeness of trial results posted at ClinicalTrials.gov and published in journals. *PLoS Med* 2013;10(12):e1001566. discussion e1001566.
- [28] Golder S, Loke YK, Wright K, Norman G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. *PLoS Med* 2016;13(9):e1002127.
- [29] Academic journals are too expensive for Harvard, Elsevier is mega greedy, and why this stinks for future librarians [Internet]. Information space. Available at <https://ischool.syr.edu/infospace/2012/05/29/academic-journals-are-too-expensive-for-harvard-elsevier-is-mega>.

- greedy-and-why-this-stinks-for-future-librarians/. Accessed October 19, 2017.
- [30] Tang E, Ravaud P, Riveros C, Perrodeau E, Dechartres A. Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles. *BMC Med* 2015;13:189.
- [31] Hartung DM, Zarin DA, Guise J-M, McDonagh M, Paynter R, Helfand M. Reporting discrepancies between the ClinicalTrials.gov results database and peer-reviewed publications. *Ann Intern Med* 2014;160:477–83.
- [32] Pradhan R, Singh S. Comparison of data on serious adverse events and mortality in Clinicaltrials.gov, corresponding journal articles, and FDA medical reviews: cross-sectional analysis. *Drug Saf* 2018; 41:849–57.
- [33] Fain KM, Rajakannan T, Tse T, Williams RJ, Zarin DA. Results reporting for trials with the same sponsor, drug, and condition in Clinicaltrials.gov and peer-reviewed publications. *JAMA Intern Med* 2018;178(7):990–2.
- [34] Zarin DA, Tse T, Williams RJ, Carr S. Trial reporting in ClinicalTrials.gov - the final rule. *N Engl J Med* 2016;375:1998–2004.
- [35] Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on trial registration 11 years after the ICMJE policy was established. *N Engl J Med* 2017;376:383–91.