



RESEARCH ARTICLE

REVISED **The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews [version 3; referees: 3 approved]**

Fala Cramond¹, Alison O'Mara-Eves ², Lee Doran-Constant³, Andrew SC Rice ¹, Malcolm Macleod ⁴, James Thomas ²

¹Pain Research, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK

²EPPI-Centre, Department of Social Science, UCL Institute of Education, University College London, London, UK

³Independent Researcher, Manchester, UK

⁴Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

v3 **First published:** 10 Dec 2018, 3:157 (<https://doi.org/10.12688/wellcomeopenres.14738.1>)

Second version: 25 Jan 2019, 3:157 (<https://doi.org/10.12688/wellcomeopenres.14738.2>)

Latest published: 07 Mar 2019, 3:157 (<https://doi.org/10.12688/wellcomeopenres.14738.3>)

Abstract

Background: The extraction of data from the reports of primary studies, on which the results of systematic reviews depend, needs to be carried out accurately. To aid reliability, it is recommended that two researchers carry out data extraction independently. The extraction of statistical data from graphs in PDF files is particularly challenging, as the process is usually completely manual, and reviewers need sometimes to revert to holding a ruler against the page to read off values: an inherently time-consuming and error-prone process.

Methods: To mitigate some of the above problems we integrated and customised two existing JavaScript libraries to create a new web-based graphical data extraction tool to assist reviewers in extracting data from graphs. This tool aims to facilitate more accurate and timely data extraction through a user interface which can be used to extract data through mouse clicks. We carried out a non-inferiority evaluation to examine its performance in comparison with participants' standard practice for extracting data from graphs in PDF documents.

Results: We found that the customised graphical data extraction tool is not inferior to users' (N=10) prior standard practice. Our study was not designed to show superiority, but suggests that, on average, participants saved around 6 minutes per graph using the new tool, accompanied by a substantial increase in accuracy.

Conclusions: Our study suggests that the incorporation of this type of tool in online systematic review software would be beneficial in facilitating the production of accurate and timely evidence synthesis to improve decision-making.

Keywords

Systematic review, automation, data extraction, graphs

Open Peer Review

Referee Status: 

Invited Referees

1 2 3

REVISED

version 3

published
07 Mar 2019






REVISED

version 2

published
25 Jan 2019

version 1

published
10 Dec 2018

Referee	1	2	3
version 3			
version 2		 report	 report
version 1	 report	 report	 report

1 **Jon Brassey** , Trip Database Ltd., UK

2 **Chris Marshall**, Newcastle University, UK

3 **Livia Puljak** , Catholic University of Croatia, Croatia

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: James Thomas (james.thomas@ucl.ac.uk)

Author roles: **Cramond F:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **O'Mara-Eves A:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Doran-Constant L:** Software, Writing – Review & Editing; **Rice AS:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Macleod M:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Thomas J:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The Wellcome Trust and Medical Research Council (MRC) supported this research through grant MR/N015665/1; and the National Institute for Health and Care Excellence (NICE) through support for Lee Doran-Constant. The views presented here are those of the authors, and not necessarily those of these institutions.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Cramond F *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Cramond F, O'Mara-Eves A, Doran-Constant L *et al.* **The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews [version 3; referees: 3 approved]** Wellcome Open Research 2019, 3:157 (<https://doi.org/10.12688/wellcomeopenres.14738.3>)

First published: 10 Dec 2018, 3:157 (<https://doi.org/10.12688/wellcomeopenres.14738.1>)

REVISED Amendments from Version 2

We would like to thank our two reviewers for their helpful feedback. While they both 'approved' the paper, they provided some useful thoughts for clarifications throughout, and we have amended the paper accordingly. We have also included [Supplementary File 6](#) and [Supplementary File 7](#) with the bibliographic details of the studies we used to identify the range of graphs included in the evaluation.

See referee reports

List of abbreviations

AUC: Area Under the Curve

CAMARADES: Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies

PDF: Portable Document Format

ROC: Receiver Operating Characteristic

SyRF: Systematic Reviews Facility

Background

Systematic review and meta-analysis are research techniques whereby as much relevant literature as can reasonably be identified on a research question is collated and analysed to give an overview of that field. In a meta-analysis, the quantitative results of the relevant research evidence are extracted from the primary research and statistically synthesised (analysed) to determine an estimate of the overall effect observed across studies and the precision associated with that effect estimate.

In order for the meta-analysis to be based on a sound dataset, the outcome data (i.e., quantitative results) need to be accurately and efficiently extracted from the primary research studies. This is often more challenging than perhaps it sounds. Studies within a review can present relevant outcome data in different ways, whether it be through providing multiple measures of the same outcome, measures at multiple timepoints, or in multiple statistical forms. These variations require skill and attention from the analyst to determine which data points need to be extracted and included in the analysis, in such a way that minimises bias and error in the selection and extraction of data. This can make the process very time consuming, even for small reviews; the labour required is obviously compounded in very large reviews, such as those seen in preclinical research.

A further complication is the actual presentation of the data, as different studies will report the outcomes in different ways, such as graphical plots, in tables, or as text. Whilst it might be difficult to aid reviewers in terms of *selecting* which pieces of data to extract through a software program, as this will inevitably vary from review to review, we considered there to be clear potential to improve both the speed and accuracy of extraction of outcome data from the included studies once the required outcomes have been identified. This report is of an evaluation of a tool designed to assist specifically with the extraction of outcome data from graphical plots, as these can be particularly time-consuming and prone to error^{1,2}. We acknowledge that there

is a variety of different ways that papers are presented (online, paper only, etc.), but we focused this work on the challenge of extracting data from papers published using the PDF file format. This format is the most ubiquitous electronic publishing medium for papers and we therefore see greatest efficiencies from improving software support in this area.

Motivation for this work

The use of systematic reviews is commonplace in clinical research, for example through Cochrane, where they are seen as the pinnacle of high-quality research synthesis, and are used frequently in clinical decision making³. In preclinical and *in vivo* fields, however, systematic reviews are less prevalent, but arguably can be just as useful, for example by guiding future research and bridging the gap between the quantity of research produced and the amount that can be effectively used by an investigator. Whilst there are some research groups pioneering the use of systematic review in preclinical fields (e.g. [CAMARADES](#)) systematic reviews have not yet gained the widespread acceptance that they have in clinical research⁴.

One of the key criticisms of systematic reviews is that, once published, they can quickly go out of date⁵. Whilst this is true for clinical systematic reviews, it is especially true for preclinical reviews due to the sheer volume and accrual rate of preclinical literature, which means that a preclinical systematic review and meta-analysis is likely to take a longer time to complete than a clinical one. For example, in a recently completed systematic review of neuropathic pain, data from 229 clinical trials required extraction⁶, whereas for the corresponding on-going preclinical systematic review data are being extracted from approximately 6000 studies. Therefore, to improve the feasibility, acceptance and usefulness of systematic reviews, methods and technologies need to be developed to speed up the process, and these advances need to be made without damaging the quality of the resultant review, and be easy and simple to disseminate on a wide scale.

Once the studies for inclusion in a systematic review have been identified, the process of 'data extraction' (or 'data collection') begins⁷. This usually involves the abstraction of data from each included study in a systematic and standardised way, from the published reports of the studies, into software from which the data can be analysed as a whole. As the synthesis of findings is conducted using these extracted data, it is vital that the data are extracted reliably. To aid reliability, data are usually extracted by two people working independently, and checked against one another. There is empirical evidence that mistakes made at this stage of the review process can affect effect estimates, and hence, review conclusions³.

Outcome data can be quite challenging to extract. Transcription errors are a common problem, with some errors not being detected until after the systematic review has been published². Moreover, some outcome data are only reported in graphs, and systematic reviewers must therefore measure values from the graphs as accurately as they can and record the results. The time taken is an important component of the cost of conducting systematic reviews and reduces their currency. While most results from clinical trials tend to be reported in tabular form,

some diagnostic test accuracy studies only report some aspects of their results in graphical form; and in the preclinical field, the reporting of results in graphs alone is commonplace.

The use of bespoke online software for conducting systematic reviews is becoming increasingly standard practice, with tools such as [Covidence](#), [DistillerSR](#), [EPPI-Reviewer](#) and [SyRF](#) offering support for a range of review types in commonly available browsers. None of these platforms support the extraction of data in graphical form, however, and reviewers need therefore to use other tools to extract data from graphs, and then transcribe this information into whichever tool they are using for synthesis. This causes two problems. First, additional effort is needed on the part of reviewers to determine which application to use, to use it consistently across the reviewing team, and then to copy data back for synthesis. Second, there are no graphical extraction tools especially written for the types of data and workflows encountered in systematic reviews. The data can therefore require some transformation before being suitable for incorporation in the review. Given these challenges, and the growing infrastructure of browser-based systematic review applications, we decided to evaluate the possibility of utilising browser technologies to improve the efficiency of data extraction from graphs. To do this, we: 1) identified relevant technologies; 2) compiled a dataset for evaluation; 3) developed a pilot user interface; and 4) undertook an evaluation of the user interface in terms of its efficiency and accuracy as compared with other extraction methods. The aim was not to create a new tool that was ready for widespread deployment, but to inform future development decisions, based on the evaluation, as to the utility of integrating such a tool in systematic review software.

Methods

Research aim

To determine if machine assistance for extracting data from a variety of graph types using a tool that has been developed and customised for systematic review needs has a meaningful impact on a) the time taken and b) the accuracy of the data extracted, compared to current methods (typically using desktop measuring software).

As there are no tools for graphical data extraction that have been developed specifically for systematic reviewers to use, we developed a pilot tool for evaluation purposes that is described below. Our primary interest, however, is in the relative performance of a tool which is designed specifically for data extraction in the context of systematic reviews (rather than this specific tool *per se*).

Research questions

1. Through a comparison of the use of our customised tool with the user's current method of outcome data extraction, is there a notable difference between the two methods in the following metrics?
 - a) Time taken
 - b) Accuracy
2. Do participants in the evaluation (with a pre-existing interest in and knowledge of systematic review) consider

this to be a viable or preferential approach compared with their current practice?

Participants and recruitment

We attempted to recruit participants from collaborators, colleagues and students on Masters-level systematic review modules using direct communication, email, social media and face-to-face interactions at conferences. The recruitment strategy targeted people who were known to have training and/or experience in conducting quantitative systematic reviews. No formal sample size calculation was performed because in this study the variation between individuals in the time taken in data extraction was not previously known, but we reasoned that a minimum of 10 participants assessing each of 23 graphs using 2 different approaches would give insights to the strengths and weaknesses of each approach, and of areas for future development.

We provided participants with an information sheet ([Supplementary File 2](#)) and consent form ([Supplementary File 3](#)), which had to be signed and returned before participation could commence. As complete datasets were most useful, participants were encouraged to complete the trial in its entirety.

Identifying graph types

To guide the development of the tool we first established the structures of graph and data typically featured in research papers, starting with the preclinical literature, where we consider the challenge of extracting data from graphs to be particularly acute. To do this we selected 34 papers ([Supplementary file 6](#)) identified in the context of systematic reviews in two different preclinical fields (animal models of neuropathic pain and animal models of D-galactose-induced aging). Papers were selected covering a range of dates to account for any changing publication patterns within the literature. These were hand-checked by F.C. to ensure that they would be relevant for our purpose (i.e. an original research paper that could be included in a review and contained outcome data presented in graphs). The number of papers required at this stage was not predetermined; instead, we continued collecting graph types until no new graph had been found for 10 consecutive papers.

Two team members that do not work in preclinical research (A.O.E. and J.T.) checked the types of graphs collated to determine whether the range of graphs in their disciplines (clinical and public health research – [Supplementary file 7](#)) were represented. The team identified that area under the curve (receiver-operator curve) plots, which are common in diagnostic test accuracy systematic reviews, were not represented, and these were added to the list of graph types.

Developing the web-based tool for graphical data extraction

We developed requirements for the graphical data extraction tool and chose a browser-based solution for ease of deployment during evaluation and because, should that evaluation prove positive, the code could be integrated within web-based systematic review software such as those mentioned above. The two main requirements were that: a) the user interface should display PDF files and support the selection of graphs from which data would be extracted; and b) the user should be able to extract data from the graphs by specifying axes values and data types,

and then by clicking appropriate points on the screen with a mouse. In terms of browser requirements, we decided that we would require HTML5 compliance, since most platforms now support this standard, and if we needed to support older browsers the cost of development would have been prohibitive.

We developed the graphical data extraction application using two existing JavaScript libraries: [PDFJS](#) (version 1.5.188) and [WebPlotDigitizer](#) (version 3.8⁸). PDFJS is a widely used library for displaying PDF files in web browsers. We used this library to display the graphs to evaluation participants and to allow them to draw a box around selected graphs. WebPlotDigitizer is a program that can extract data from graphs that are uploaded in a PNG or JPEG format, so we used JavaScript to 'send' the graph image to WebPlotDigitizer, and this library was customised to support our workflow and the data types common in systematic reviews. We now describe in more detail the workflow in the customised tool, and the software development undertaken to support it.

Our starting point for the design of this workflow was that users would be extracting data from a PDF file and, as part of this process, would encounter graphs within the file from which they would need to extract numeric data. We simulated this initial point of entry by using PDFJS to display PDF files. This tool can display the majority of PDF files, and we encountered no problems when using it. We modified the default display of the tool in two ways for the purposes of this evaluation. First, we suppressed the appearance of buttons and functionality that were unnecessary for our purposes (e.g. the ability to 'zoom' into / out of the page). Second, we wrote functionality that enabled users to draw a box around content in the PDF with a mouse – in our use scenario, users were expected to draw a box around a graph – and extract this part of the PDF file as an image to be used in the next part of the workflow. [Figure 2](#) shows the display of graphs in the PDFJS tool with a 'box' drawn around the graphic. The user then clicks the box to move to the next part of the workflow where we incorporated the WebPlotDigitizer tool⁸. WebPlotDigitizer is an online tool which supports the extraction of numeric data from many types of graphs. We customised the tool so that it supported both the data types that systematic reviewers use explicitly, and also our expected workflow.

After drawing a box around their selected graph and clicking the graph, the user moves into the user interface of WebPlotDigitizer⁸. Our modifications to WebPlotDigitizer fell into three main areas which are outlined in detail below: 1) user selection of specific data types and structures at the outset; 2) the addition of an interactive data table, which captures the data in a structure which is suitable for use in subsequent meta-analyses; and 3) data export.

We modified the normal point of entry to WebPlotDigitizer to display a menu of data types, series and data points for users to specify exactly what type of graph they were going to extract data from ([Figure 3](#)). In the example screenshot in [Figure 3](#), we can see that the graph has three data points for each of four series, and each data point has a mean and confidence intervals around

it. The user specifies this information, as well as whether the graph has one or two axes (in the example, there is only one axis: the y-axis). The tool then utilises the standard functionality of WebPlotDigitizer which supports the calibration of axes. This involves clicking the mouse on the origin (i.e. the bottom left hand side of the graph in the example) and the highest value in the axis (in the example, on the 10). The user then enters the values and the software can calculate the correct values for the positions of any mouse-clicks in between the two.

After specifying the position of the axis (or axes) and calibrating them, the user then enters the main data extraction screen ([Figure 4](#)). Here, our development work focused on a new interactive data table which can be dragged to any part of the screen, or docked on the far right-hand-side ([Figure 5](#)). The structure of the data table matches the parameters entered previously; in the example, we have three columns for data: the mean, and the upper, and lower confidence intervals. This data structure is multiplied by the number of series and the number of data points that have been specified by the user. The titles of the series are editable text boxes and the user can use the tab key to move between them, entering series titles.

The user then clicks in the cell that they want to enter data into and can then use the mouse to click data points on the graph. The user interface automatically advances between cells. For example, in the data table shown in [Figure 5](#), the user would click in the top cell in the 'mean' column and click on the first data point for 'Model + vehicle'. This value then appears in the relevant cell. The 'active' cell automatically advances to the upper confidence interval, which is filled in when they click on the relevant point on the graph. Note that the data table is 'aware' of the various data types that it contains and calculates this value as a difference between the mean and the corresponding value on the y-axis, rather than just the y-axis value. The data table can have radically different structures; for example, it can capture both individual and aggregate level data when necessary ([Figure 6](#)).

The 'point-and-click' interface, working alongside the data table means that the user is able to extract the relevant numeric data from a graph in a matter of seconds. Moreover, in order to assist with accurate mouse positioning, WebPlotDigitizer⁸ contains by default a window on the top right which shows a 'zoomed in' display of the current mouse position.

The final piece of development that we undertook was to save the data back up to the server ready for analysis.

Development took place iteratively during November 2016 to June 2017 by a highly experienced JavaScript developer (LD-C) resulting in the customised tool which was placed online for evaluation in June / July 2017. The integration and customisation of the two existing JavaScript libraries, [PDFJS](#) (version 1.5.188) and [WebPlotDigitizer](#) (version 3.8⁸), resulted in a prototype workflow designed specifically around the needs of systematic reviewers. The aim was not to create a new tool that was ready for widespread deployment, but to inform future development decisions, based on the evaluation, as to the utility of integrating such a tool in systematic review software.

Evaluation design

We used a non-inferiority trial design to evaluate the graphical data extraction application, with each participant extracting data from graphs using their current methods of data extraction and the new, customised graphical data extraction application. The study was approved by the UCL Institute of Education Research Ethics Board (reference REC 944.)

Our primary aims were to determine whether there were differences in time taken and accuracy between a user's current approach and the new approach to data extraction. We also sought feedback from users as to the usefulness of the new, customised tool.

We identified 5 broad classes of graph and created 23 examples (5 bar, 5 line, 5 scatter, 3 dot plot, and 5 box and whisker) (Supplementary File 5) in SigmaPlot version 10 using fictitious data and expressed to 3 significant figures, so that the true value for each data point was known; the 'new tool' condition had an additional class of graph, ROC/AUC, for which 4 graphs were created. Participants were required to extract data from graphs using both their current methods of data extraction and the new graphical data extraction application. For each method of extraction they worked through all 23 graphs in the same order (plus the 4 AUC/ROC graphs in the 'new tool' condition, and whether

they started with the new method or their current method (defined as their preferred method that is used most often when extracting data) was determined at random by software code embedded within the study website.

Current methods condition

The evaluation aimed to compare the purpose-built workflow, described above, with current practice. Because 'current practice' varies from person to person, we did not specify exactly which method participants should use, as we wanted them to use the one that they would naturally use – whether that was a specific tool, or simply holding a ruler against the computer monitor. Participants were therefore instructed to extract data from plots in this condition using whatever methods they typically currently use. Participants reported that the following tools were used in this condition: Universal Desktop Ruler (3), In-built Adobe Acrobat measuring tool (3), WebPlotDigitizer (2), and Grabit (1). The Qualtrics survey platform was used to collect data for this condition, whereby the graphs were uploaded alongside a table where the participant was asked to record their extractions. This software allows an accurate timing per graph to be collected. Figure 1 is a screenshot of one of the graphs with the table for entry of extracted data shown below. A copy of the platform as presented to participants can be found at https://imperial.eu.qualtrics.com/jfe/form/SV_eXnjY1YyPSY1mDj.

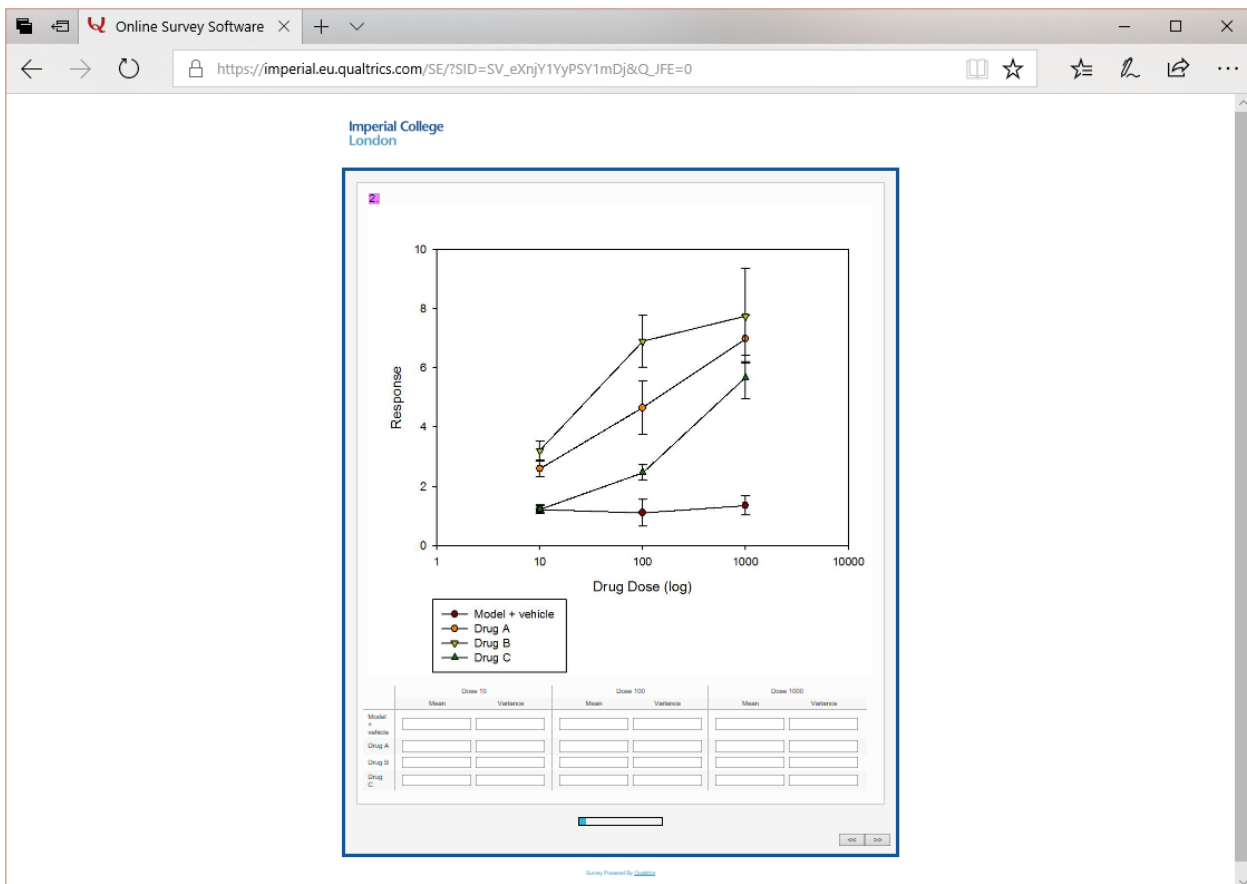


Figure 1. The 'current methods' data collection tool in the Qualtrics platform.

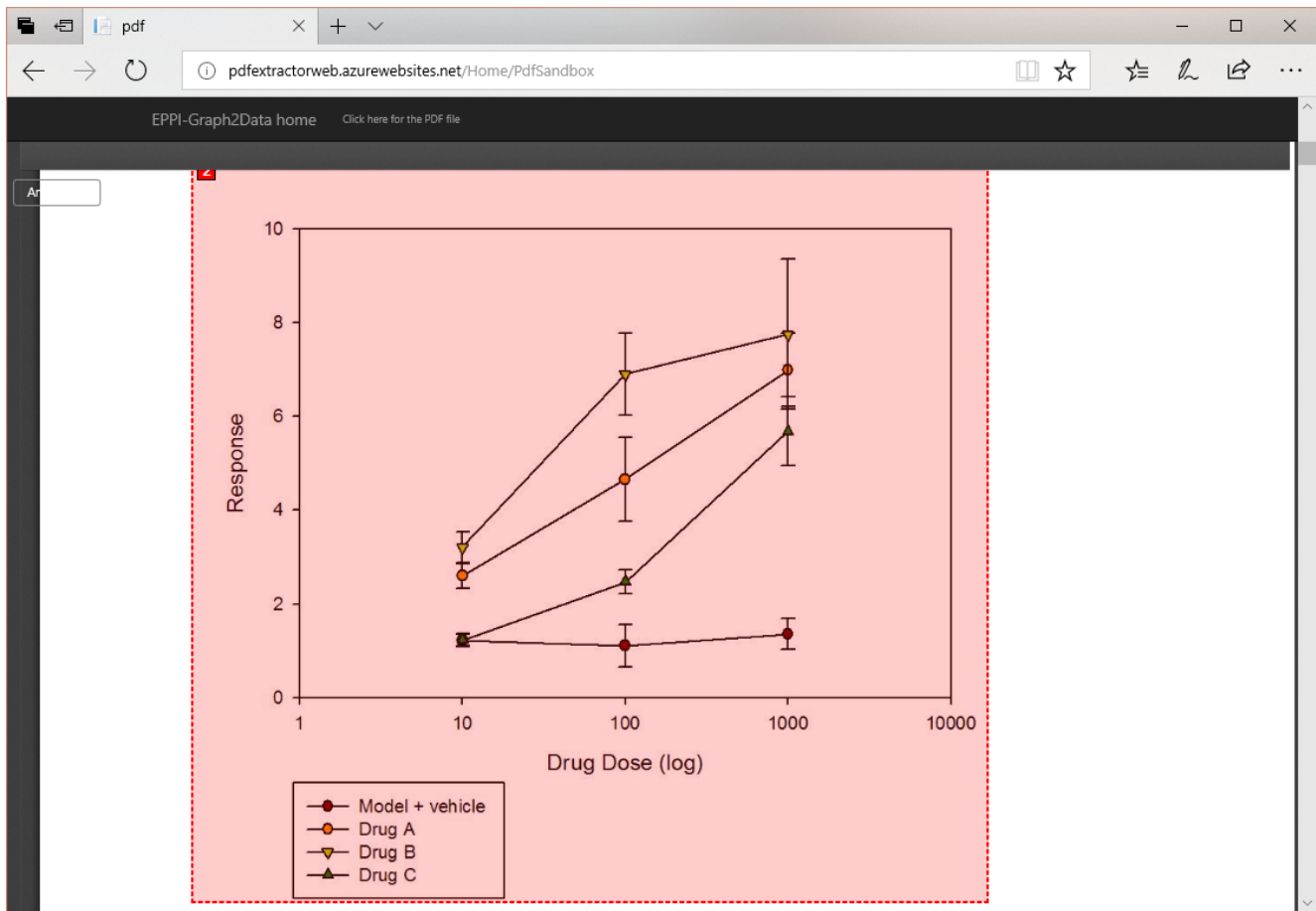


Figure 2. The graph as displayed in PDF.JS.

Because of the challenges in manual data extraction from ROC/AUC graphs, these were not offered in this set.

New graphical data extraction application condition

The graphs for the new graphical data extraction application were the same and in the same order as the current methods condition, with the addition of 4 AUC/ROC graphs.

The evaluation website for the graphical data extraction application was hosted at: <http://pdfextractorweb.azurewebsites.net/>. As well as supporting the data extraction problem itself, the graphical data extraction application measured the time that participants spent extracting data from each graph automatically. Participants were given comprehensive instructions, including a YouTube tutorial video (https://www.youtube.com/watch?v=tzg-NUV-wcg&feature=em-upload_owner) and instructed not to leave the platform running when not in use, as this would affect the accuracy of the time measurements. Data validity were also part of the subsequent analysis

Participant experience

We used a qualitative survey hosted on the surveymonkey.com platform. The questions focused on the background experience

of the participants and their perceptions on the ease, speed, and features of the tool. Participants were also asked to indicate their preferred method for future extractions and were able to submit suggestions for development of the tool. This was filled in after completion of the trial. The questions on the survey are presented in [Supplementary File 1: https://www.surveymonkey.co.uk/r/G5XYQDS](https://www.surveymonkey.co.uk/r/G5XYQDS).

Quantitative analysis

We used a non-inferiority trial design to seek to demonstrate that the novel process (the use of a graphical data extraction application) was not meaningfully worse than the existing process (current methods of data extraction). Data were analysed in Microsoft Excel. The analysis process is outlined below.

To establish the time taken to extract the data for each method we used a within-subjects design. As participants were required to extract the same data from the same graphs in each condition, it was possible to directly compare how long it took using each method of extraction. To measure differences between approaches we calculated by subtraction, for each graph and for each participant, the difference in time taken between each approach, such that a positive value would indicate that the

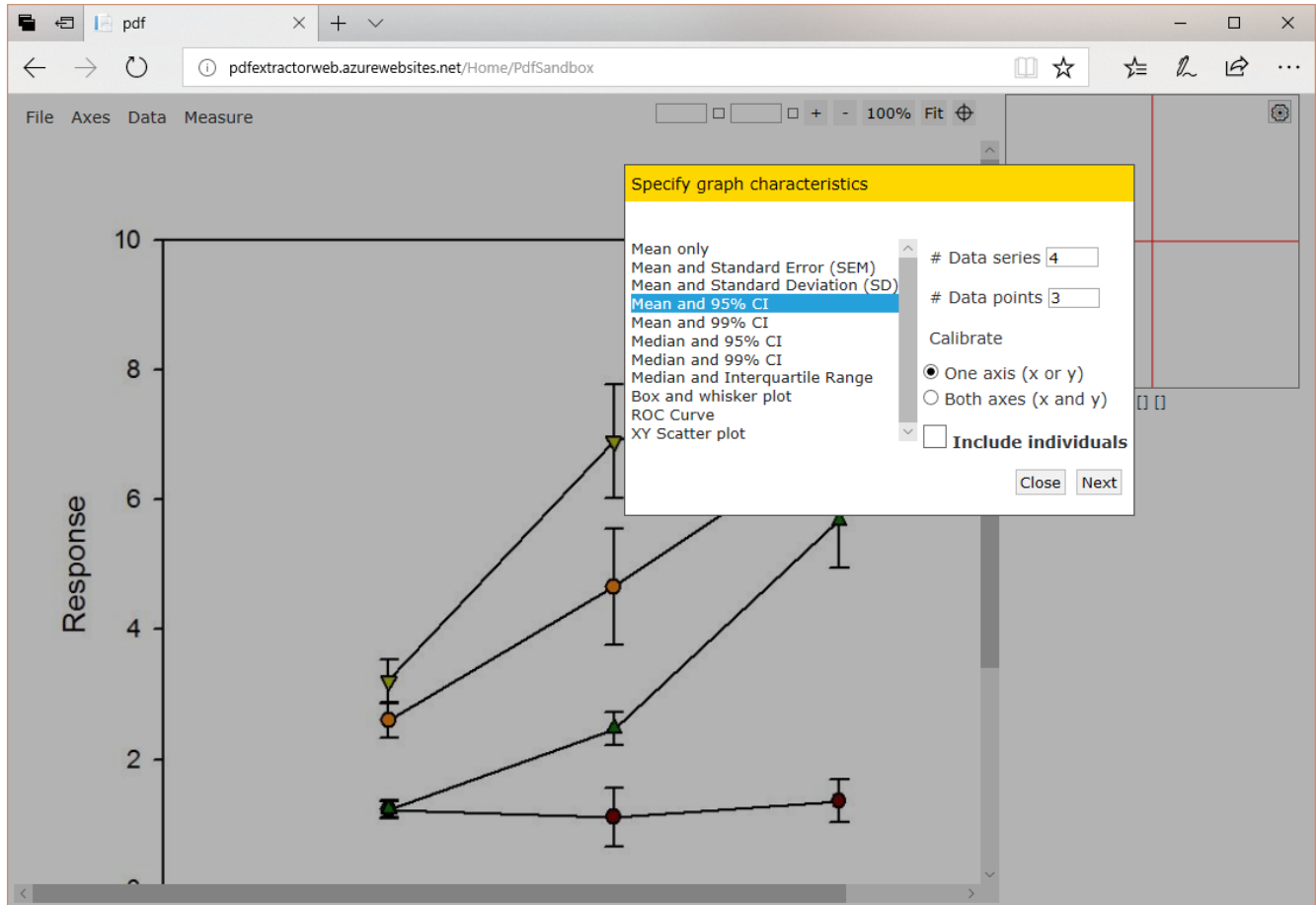


Figure 3. Specification of graph characteristics.

current methods took longer than the new method. Then, for each graph we calculated a mean difference in time taken across participants, along with the standard deviation; and we also calculated the total time taken for all 23 graphs represented in both conditions, and expressed this as minutes.

Note that analysis of the difference in time taken for the two conditions could not be computed for the four AUC graphs because they were only presented in the new graphical data extraction application condition (i.e., we do not have data for the four AUC graphs in the current methods condition).

To establish the accuracy of data extraction, we compared extracted values with the known true values used to render the graphs.

We first defined the tolerable bounds of an ‘accurate’ extraction for each graph (Supplementary File 4). We calculated the bounds as $1/20^{\text{th}}$ of one increment in the scale of the graph outcome axis (usually the y-axis). For example, if the outcome axis scale had increments of 10, then a bound of ± 0.5 around the true data point was set. If a given true data point had the value of 6, with a tolerable bound of ± 0.5 , then we would accept any value between 5.5 and 6.5 as accurate for that data point. The

bounds for each graph are shown in Supplementary File 4. Extracted data points lying on or within these bounds were considered accurate, while those above the upper bound or below the lower bound were considered inaccurate.

In a real systematic review, data extraction is usually performed by two individuals working independently because 100% accuracy in data extraction cannot be guaranteed; errors of one extractor can be detected when disagreement is observed with the other extractor, and these data points identified for third person reconciliation. For each data point, we determined whether 80% or more participant responses were within the tolerable bound. For each graph, we were then able to determine what proportion of data points were ascertained with sufficient accuracy.

To give a summary estimate of differences in the accuracy of data extraction using the different methods, we calculated the difference between the percentage of accurate data points using the new method and that using conventional methods. We determined in advance that we would consider that the new method was inferior to current methods if the point estimate of sufficient accuracy was greater than or equal to 5% lower than current methods (i.e., the new, customised tool would be considered inferior if $\text{SufficientCurrentMethod} - \text{SufficientNewMethod} \geq 5\%$).

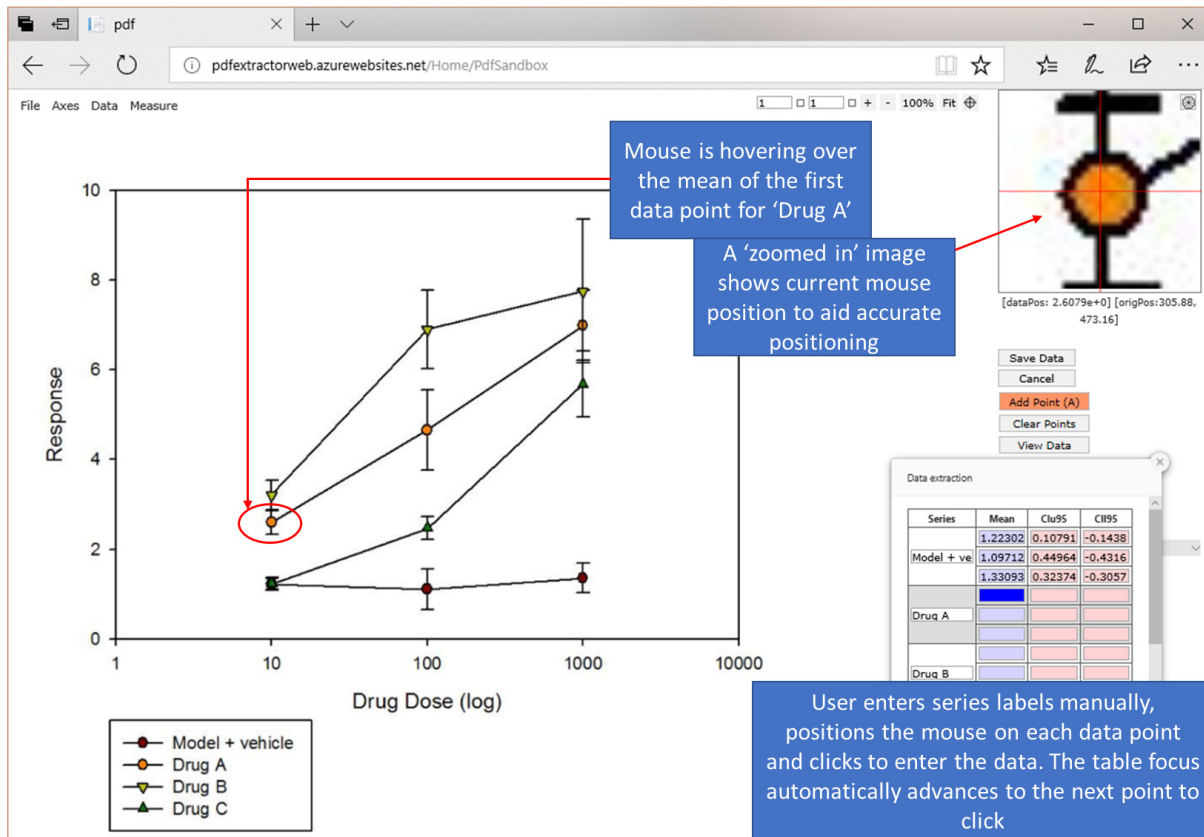


Figure 4. Data extraction.

Under such circumstances, substantial redesign of our approach would be required.

We also calculated an odds ratio for obtaining a sufficiently accurate data point in the new method compared to the current method as: $(\text{SufficientNewMethod} / \text{InsufficientNewMethod}) / (\text{SufficientCurrentMethod} / \text{InsufficientcurrentMethod})$, where the values represent the number of data points that were of sufficient (or insufficient) accuracy in the two conditions (new and current methods).

Qualitative analysis plan

A secondary aim of the project was to consider users' reactions to the new, customised tool. Analysis of the multiple-choice questions involved examination of frequencies and percentages of participant responses. Analysis of the open-ended text responses involved coding the text into categories (themes) that were derived from the data (i.e., not *a priori*); for example, free text comments about how quickly the participant extracted data were coded as relating to the theme of 'speed'. The frequencies of themes mentioned across participants were examined. To protect the anonymity of the participants and encourage completion, the survey data were not linked to the responses from the data extraction conditions.

Results

Recruitment

Emails were directly sent by a members of the research team to more than 50 people. We are unable to state how many people were exposed to the social media adverts, and therefore cannot provide an accurate number of how many people were indirectly approached.

A total of 32 consent forms were returned. Of these individuals, 10 completed the trial, 9 never started the trial, 7 partially completed the trial and 6 were excluded or dropped out. Recruitment commenced 30/06/17 and was completed 01/10/17. Data for a total of 10 participants were included in the analyses. The relatively high drop-out rate can be explained by the time demands of the evaluation. The evaluation – and especially the 'current methods' component took some participants several hours to complete. This was necessary in order to collect sufficient data to be able to compare the two approaches across so many different types of graph, but it did affect recruitment and retention.

Time

As described in the methods, we calculated the difference in times as the time for the current methods condition minus the

Data extraction

Series	Mean	Clu95	CI195
	1.20748	0.17006	-0.1190
Model + vehi	1.10544	0.45918	-0.4591
	1.36054	0.34013	-0.3231
	2.56802	0.28911	-0.2210
Drug A	4.65986	0.88435	-0.8843
	6.97278	0.81632	-0.5952
	3.16326	0.39115	-0.2891
Drug B	6.90476	0.83333	

Figure 5. Data table.

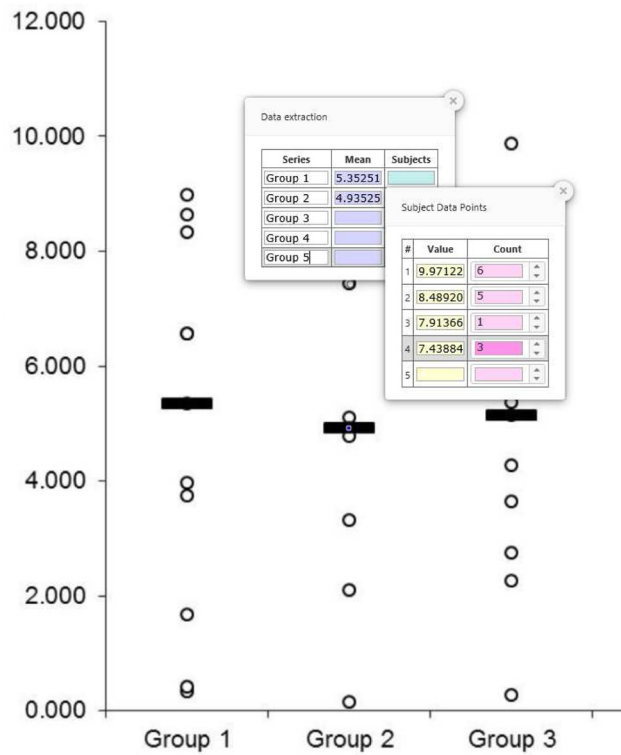


Figure 6. An alternative type of data table, supporting individual and aggregate level data.

time for new graphical data extraction tool within a participant, so that a positive value would indicate that the current methods took longer than the new method. The mean of these differences across participants was calculated to give (Xg^-) (in seconds); the results of which are reported for each graph in Table 1.

For each graph, the average time taken was less when using the new graphical data extraction tool compared with the usual approach used by participants, with some differences of more than 10 minutes. Overall, the mean time taken to extract data was 352 s (5 min 52 s) less using the new, customised tool than using

Table 1. Mean and standard deviation of the time difference for each graph across n participants.

Graph number	Mean (\bar{X}_g) time difference, s	Standard deviation	Participants, n
1	180.02	240.55	10
2	363.60	255.27	10
3	314.59	225.16	9
4	140.22	121.24	10
5	113.68	132.87	10
6	486.54	298.17	10
7	463.91	410.30	10
8	167.77	153.72	9
9	332.57	252.34	8
10	546.28	649.84	7
11	412.55	243.81	9
12	564.25	820.52	8
13	210.15	169.28	8
14	377.24	466.44	8
15	281.76	331.66	8
16	478.74	404.05	8
17	691.20	738.84	7
18	119.62	104.42	8
19	93.34	151.23	9
20	650.31	750.77	9
21	469.19	804.92	8
22	373.28	462.56	9
23	270.24	258.40	8

Note: A positive time difference indicates that the current methods condition took longer than the new graphical data extraction application method condition.

the conventional approach (median, 364 s; IQR, 180–469 s; range, 93–691 s).

Accuracy

As described in the Methods, we considered whether a given data point was sufficiently accurate if at least 80% of participants' responses fell within a tolerable boundary around the true value. The number of data points that were of sufficient accuracy or insufficient accuracy were summed for each graph. The results for each graph, presented by condition, are shown in Table 2. Recall that the new tool would be considered inferior if $\text{SufficientCurrentMethod} - \text{SufficientNewMethod} \geq 5\%$. Overall, the current method ascertained data with sufficient accuracy for 41% of data points, compared with 70% for the new approach, for a difference of -29%, which is substantially better than our prespecified non-inferiority value of 5%.

(Here, anything less than 5% difference is favourable to the new method). The odds ratio of getting a sufficiently accurate data point compared to an insufficient data point in the new method compared to current methods was 3.34 (95%CI = 2.51, 4.44).

Survey results

A total of nine participants completed the qualitative survey. They were employed at a higher education institute (n=3), by a governmental agency (n=2) or were students (n=3 doctoral and n=1 masters). Their disciplines were preclinical science (n=4), statistics (n=1), clinical science/medicine (n=2) and social sciences (n=2). All had performed at least one stage of a systematic review and seven stated they had extracted outcome data previously. Tools previously used for extracting data from graphs included the universal desktop ruler (n=3 participants), Adobe measuring tool (n=4 participants), Web Plot Digitizer (n=3 participants) and Excel Grabit (n=1 participant). Three participants stated they had not previously extracted graphical data. Unfortunately, because the survey and trial data were not linked, we could not explore whether the background or experiences of the participants' might have been associated with their performance in the trial.

The percentage of respondents that either 'agreed' or 'strongly agreed' with statements evaluating their satisfaction with the features of the new graphical data extraction tool are depicted in Figure 7. They show strong support for the tool as compared with other methods, although these and subsequent answers suggest that additional development may be needed. Raw data is available on Zenodo⁹.

All respondents indicated that if they had to extract a third set of similar graphs using just one of the methods they would choose the new online tool. In a free text box they were asked why this selection was made. Comments referred to speed (n=7), accuracy (n=4), and ease of use (n=5).

Lastly, participants had an option to submit suggestions for improvement of the tool; these included bug-fixing, an undo button, functionality of plotting the points, and an interface to allow the tool to interact with a data storage tool.

Discussion

Summary of findings

We have shown that our new graphical data extraction tool¹⁰ is not inferior to users' preferred current approaches. Our study was not designed to show superiority, but suggests that, on average, participants saved around 6 minutes per graph using the new tool, accompanied by a substantial increase in accuracy. Indeed, that gain in accuracy is likely to be accompanied by further time-saving, as the number of outcome measures identified for reconciliation by a third reviewer will fall as a consequence. If our findings are confirmed, this would have profound implications for the conduct of systematic reviews where extraction of data from graphs is required. Our tool also received positive feedback from users in terms of its ease of use, fitness for purpose and perceived efficiency.

Table 2. Frequency per graph of data points deemed sufficient accuracy or insufficient accuracy, with percentage of data points that are sufficient accuracy, by condition.

Graph	Current methods condition			New graphical data extraction application condition		
	Sufficient accuracy	Insufficient accuracy	Percent sufficient data points	Sufficient accuracy	Insufficient accuracy	Percent sufficient data points
1	3	1	75.00%	4	0	100.00%
2	16	8	66.67%	18	2	90.00%
3	7	5	58.33%	7	5	58.33%
4	1	9	10.00%	10	0	100.00%
5	8	4	66.67%	6	0	100.00%
6	15	33	31.25%	17	0	100.00%
7	14	6	70.00%	5	15	25.00%
8	9	11	45.00%	9	11	45.00%
9	16	16	50.00%	18	0	100.00%
10	could not match data so removed from analysis					
11	0	30	0.00%	0	20	0.00%
12	3	33	8.33%	20	14	58.82%
13	10	10	50.00%	20	0	100.00%
14	37	3	92.50%	14	0	100.00%
15	could not match data so removed from analysis					
16	12	28	30.00%	10	12	45.45%
17	22	33	40.00%	8	12	40.00%
18	0	12	0.00%	5	5	50.00%
19	10	2	83.33%	12	0	100.00%
20	could not match data so removed from analysis					
21	22	38	36.67%	27	9	75.00%
22	12	30	28.57%	14	0	100.00%
23	10	14	41.67%	20	0	100.00%
Totals	227	326	41.05%^a	244	105	69.91%^a

Notes: Three of the graphs (10, 15, 20) had incompatible data because participants in the new graphical data extraction application condition selected too many different data input types, so a comparison could not be made. The total number of data points in the two conditions differs due to issues including missing data or incorrect selection of graph type in the new graphical data extraction application condition. ^aThis value represents the mean for this column, not the total.

Evidence of feasibility of further development and dissemination

For a new technology to be worth developing and disseminating, at least two conditions need to be in place. Firstly, the technology must be not inferior to existing tools. Secondly, the technology must be seen by the end users as preferable to existing tools. We believe that this study provides sufficient evidence that these two conditions have been met.

The potential cost- and time-saving aspects of the graphical data extraction tool are likely to be substantial. The results showed a mean reduction of nearly 6 minutes in time taken to extract data from graphs compared to existing methods, which could translate to a substantial time saving per systematic review

publication, due to reduced reviewer time. In practice, this time saving would be amplified, as it is advised that data in systematic reviews should be extracted by a minimum of two reviewers to reduce errors¹ and potentially even a third reviewer to resolve discrepancies.

Furthermore, as the graphical data extraction tool showed a considerable improvement in accuracy; this will also decrease time as the third reviewer will have fewer discrepancies to resolve.

Aside from the time-saving aspect, the improvement in accuracy alone is compelling evidence for the further development of the software, as it ultimately may lead to more precise systematic reviews and meta-analyses. In line with Jelacic Kadic *et al.*,

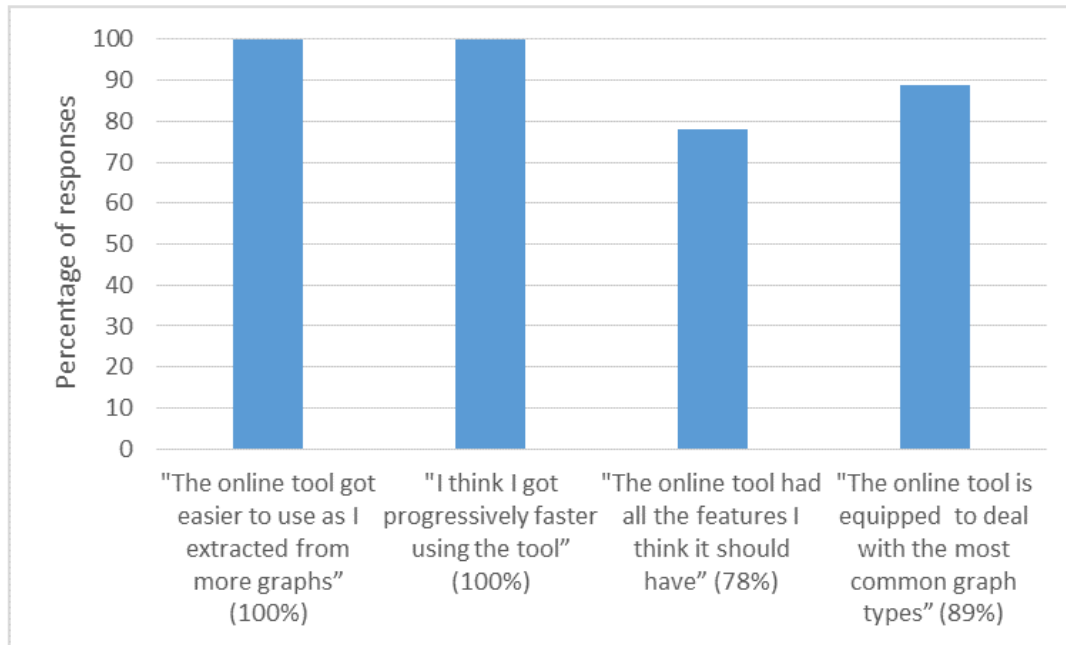


Figure 7. Satisfaction with the features of the new graphical data extraction tool: percentage of respondents who ‘agreed’ or ‘strongly agreed’.

who evaluated an electronic data extraction tool with a paper-based method, we found that the use of a graph extraction tool leads to more accurate data extraction¹¹.

We note that the few graphs for which graphical data extraction application had very poor performance were cases in which some participants had selected completely the wrong graph type; this means that our estimates for the accuracy of data extraction from graphs for the new graphical data extraction application condition are considerably below that which is probably likely in real life conditions. It also suggests that some training or further guidance on graph type selection within the tool (as depicted in [Figure 3](#)) is required.

Lastly, the qualitative survey provides evidence that reviewers prefer the new data extraction tool, described hereby, to several desktop electronic methods of data extraction that are currently in use. This suggests that the tool will be acceptable and credible to the proposed users, which is necessary for its uptake.

Ultimately, there is strong evidence from the trial of the graphical data extraction tool that the further development and dissemination of this technology is worthwhile. The initial costs of implementation, training, and monitoring, would be offset by the impact of widespread use, leading to increased output of accurate systematic reviews, especially in preclinical topics where a large proportion of the outcome data are extracted from graphs.

Future work

As it currently stands, the technology developed here has limited ‘real life’ use. For it to become a useful part of the systematic review process it would need integration with other platforms used to facilitate systematic review and meta-analysis. An example is the [SyRF platform](#) (CAMARADES) which allows for screening and annotations for risk of bias using technology developed in other work packages for preclinical studies. Another example is [EPPI-Reviewer](#)¹², a tool widely used in clinical and social scientific evidence synthesis, which is the core evidence synthesis platform for the UK’s National Institute for Health and Care Excellence. The new online tool will be integrated within these two platforms and, since it is open source, it is available for integration within other systematic review platforms too.

The ultimate aim for the future would be “living” systematic reviews, which are updated constantly as new research evidence becomes available¹³. Given the scarcity and expense of human input, the use of new technologies—including automation—is being evaluated for these types of reviews¹⁴. Moreover, the human/machine axis may not be considered as binary opposites, as citizen science platforms, such as [Cochrane Crowd](#), have shown that workflows can be developed that maximise the efficacy of human and machine contribution.

Unfortunately, the complete automation of outcome data extraction from graphs currently seems unlikely due to the varied nature of graphs and, as in most reviews, not every graph requires extraction, so human intelligence is required to decide

which graph is the most relevant. However, for us to move towards goals of minimal human time to get maximum output, specifically for outcome measure extraction, we propose that further software development work be undertaken to support the automatic:

- identification of graph axes and their values, and optical character recognition to digitise text (e.g. axes labels), so a reviewer does not need to enter these manually¹⁵
- recognition of figures that are potentially relevant for a research question
- recognition of figures that are definitely not relevant for a research question
- flagging of discrepancies between reviewers and identification of patterns within these, so that time is saved when resolving discrepancies.

Limitations

There are several potential limitations to this evaluation. First, we are unlikely to have identified all graph types that are present in the clinical and preclinical literature. However, we believe that the graph types identified include most commonly used formats; other formats such as flow cytometry outputs are rarely extracted in the context of meta-analysis and so their omission is unlikely to have a major impact on our findings. This is supported by the observation that 89% of trial participants either agreed or strongly agreed that the online tool covered the most important graph types.

Second, this is a small study. We did not set out to show the superiority of the new, customised tool, and no conclusions of superiority should be drawn. However, we believe that it is reasonable to characterise the effectiveness of the tool as being promising.

Third, the extent to which the trial accurately reflected ‘real-life’ data extraction might be questioned, because in real-life, the reviewer would also be reading the rest of the paper, and maybe only extracting one time point from each graph and extracting other information such as group numbers or details of the paper. However, this trial aimed to separate the data extraction from this, so it could be analysed as a separate entity without other confounding aspects.

Finally, although not explicitly measured here, we observed that most data points that were extracted with sufficient accuracy using the graphical data extraction application had 100%

of responses within the tolerable bounds; whereas in the current methods, even those that achieved sufficient accuracy often had responses outside of the tolerable bounds. Had we explored accuracy at the individual participant level, we would have likely seen even greater gains in accuracy in the graphical data extraction application condition.

Conclusions

We have detailed here the motivation for, and development of, the customisation of two existing JavaScript libraries to create a new web browser-based tool to facilitate the extraction of quantitative data from graphs embedded in pdf files. We evaluated its utility in terms of its efficiency and accuracy, finding that it demonstrated non-inferiority compared to current practice in both dimensions. Our study suggests that the incorporation of this type of tool in online systematic review software would be beneficial in facilitating the production of accurate and timely evidence synthesis to improve decision-making.

Data availability

Raw data associated with this study, including results of the survey, are available on Zenodo, DOI: <https://doi.org/10.5281/zenodo.1482487>.

Software availability

Source code available from: <https://github.com/EPPI-Centre/Graph2Data>.

Archived code at time of publication: <https://doi.org/10.5281/zenodo.1484506>.

License: [GNU Affero General Public License](#)

Grant information

The Wellcome Trust and Medical Research Council (MRC) supported this research through grant MR/N015665/1; and the National Institute for Health and Care Excellence (NICE) through support for Lee Doran-Constant. The views presented here are those of the authors, and not necessarily those of these institutions.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to acknowledge with thanks the support we have received from our funders for this work; the time and effort given by participants in the evaluation; and the developers of PDF.JS and WebPlotDigitizer.

Supplementary material

Supplementary File 1. Survey questions.

[Click here to access the data](#)

Supplementary File 2. Information sheet given to each participant during recruitment.[Click here to access the data](#)**Supplementary File 3. Consent form for subjects taking part in the study.**[Click here to access the data](#)**Supplementary File 4. The tolerable bounds for determining accuracy of each data point, by graph.**[Click here to access the data](#)**Supplementary File 5. The graphs used for evaluation of the methods of data extraction.**[Click here to access the data](#)**Supplementary File 6. List of 34 pre-clinical studies.**[Click here to access the data](#)**Supplementary File 7. List of 12 public health studies.**[Click here to access the data](#)**References**

1. Buscemi N, Hartling L, Vandermeer B, *et al.*: **Single data extraction generated more errors than double data extraction in systematic reviews.** *J Clin Epidemiol.* 2006; **59**(7): 697–703.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Mathes T, Klaffen P, Pieper D: **Frequency of data extraction errors and methods to increase data extraction quality: a methodological review.** *BMC Med Res Methodol.* 2017; **17**(1): 152.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. National Institute for Health and Care Excellence: **Developing NICE Guidelines: The Manual [Internet].** London: National Institute for Health and Care Excellence (NICE); 2015 Jul 22. Process and Methods Guides No. 20.
[PubMed Abstract](#)
4. Williams A, Sharpe S, Verreck F, *et al.*: **Response to: Systematic review: animal studies of TB vaccines.** *Int J Epidemiol.* 2016; **45**(2): 583–584.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Shojania KG, Sampson M, Ansari MT, *et al.*: **How quickly do systematic reviews go out of date? A survival analysis.** *Ann Intern Med.* 2007; **147**(4): 224–33.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Finnerup NB, Attal N, Haroutounian S, *et al.*: **Pharmacotherapy for neuropathic pain in adults: a systematic review and meta-analysis.** *Lancet Neurol.* 2015; **14**(2): 162–173.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Gough D, Oliver S, Thomas J: **An Introduction to Systematic Reviews (2nd Edition).** London: Sage, 2017.
[Reference Source](#)
8. Rohatgi A: **WebPlotDigitizer.** Austin, Texas, USA, 2018.
[Reference Source](#)
9. Cramond, O'Mara-Eves, Doran-Constant, *et al.*: **The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews (data repository) [Data set].** *Zenodo.* 2018.
<http://www.doi.org/10.5281/zenodo.1482487>
10. Crammond F, O'Mara-Eves A, Doran-Constant L, *et al.*: **The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews (Graph2Data tool) (Version 0.1).** *Zenodo.* 2018.
<http://www.doi.org/10.5281/zenodo.1484506>
11. Jelacic Kadic A, Vucic K, Dosenovic S, *et al.*: **Extracting data from figures with software was faster, with higher interrater reliability than manual extraction.** *J Clin Epidemiol.* 2016; **74**: 119–123.
[PubMed Abstract](#) | [Publisher Full Text](#)
12. Thomas J, Brunton J, Graziosi S: **EPPI-Reviewer: software for research synthesis.** 2010.
[Reference Source](#)
13. Elliott JH, Turner T, Clavisi O, *et al.*: **Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap.** *PLoS Med.* 2014; **11**(2): e1001603.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Thomas J, Noel-Storr A, Marshall I, *et al.*: **Living systematic reviews: 2. Combining human and machine effort.** *J Clin Epidemiol.* 2017; **91**: 31–37.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Tsafnat G, Glasziou P, Choong MK, *et al.*: **Systematic review automation technologies.** *Syst Rev.* 2014; **3**: 74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Referee Status:



Version 2

Referee Report 14 February 2019

<https://doi.org/10.21956/wellcomeopenres.16420.r34676>



Chris Marshall

Institute of Health and Society, Newcastle University, Newcastle, UK

Thank you to the authors for addressing my initial comments so comprehensively. The extra detail and clarity concerning the aims, objectives and methods is excellent... I get it now! My only other minor comment concerns a couple of lines in the final paragraph of Page 6:

"a prototype workflow designed specifically around the needs of systematic reviewers"

"The aim was not create a new tool that was ready for widespread deployment but to inform future development decisions, based on the evaluation, as to the utility of integrating such a tool in systematic review software"

My comment is really just to say that I think these bits are great and cut right to the heart of the paper and what the work aimed to achieve. Perhaps these lines could be brought forward or integrated into the 'Research aim' section at the beginning of the methods?

Valuable paper in this space I think, as we need more evaluations of systematic review tools!

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 02 Mar 2019

James Thomas, University College London, UK

We would like to thank Chris Marshall for his second review, and are pleased that our revisions have addressed his concerns. We have inserted "The aim was not create a new tool that was ready for widespread deployment but to inform future development decisions, based on the evaluation, as to the utility of integrating such a tool in systematic review software" into the 'aims' section as suggested, to ensure our evaluation focus is clear at the start. We have also made a number of other amendments to improve the paper's clarity in response to our other reviewer.

Competing Interests: No competing interests were disclosed.

Referee Report 11 February 2019

<https://doi.org/10.21956/wellcomeopenres.16420.r34678>



Livia Puljak 

Cochrane Croatia, Catholic University of Croatia, Zagreb, Croatia

After reviewing the manuscript 'The development and evaluation of an online application to assist in the extraction of data from graphs for use in systematic reviews', here is my opinion: this is a relevant manuscript, as the process of evidence synthesis would benefit from further methodology refinements, and certainly new ideas for making systematic review methods faster and more accurate are very welcome.

I would like to suggest the following minor revisions:

Abstract

- The extraction of statistical data from graphs in PDF files is: here, I would remove 'in PDF files' because some manuscripts do not have online edition so one would need to rely on printed version. Also, some older manuscripts that are online do not have a PDF version for download. Now, this can be removed if the tool can accept other files too. But if not, it should be specified that the tool in this version was developed only to work with pdf files.
- Please specify clearly in the abstract what is the 'comparison to standard practice'. I am not sure that there is a standard practice here in a general sense. Many researchers simply disregard data from figures as a 'practice'. Even in some Cochrane reviews one can find a note that data presented only as graphs will be excluded.
- Later in the text I found more detailed description about what was considered a 'standard practice' and I noticed that four electronic data extraction methods were used, and that ruler-paper method was not used. I think this should be clearly specified in the abstract; that the new workflow was compared with other electronic solutions for data extraction that the participants were using most commonly in their research work.
- It would be great to mention already in the abstract how many participants took part in the evaluation of design. I would also mention in the abstract that you had qualitative survey with participants. These are all relevant information to present your study more comprehensively.
- Explain what does it mean 'users' prior preferred current approaches' and how does it compare to 'standard practice' from the previous paragraph of the introduction. See comment above. These terms in abstract should be clearly specified.
- This statement is very unhelpful 'there may be a saving in time of around 6 minutes per graph,' because there are different types of graphs and savings of time will not be the same if one graph has ten data points to extract, or twenty, or only three, for example. It would be better to be very clear about where is the time savings – how much time is saved to extract one data point if the authors really want to present time saved. This information about six minutes per graph needs to be removed and revised. For example, you can talk about savings per one data point, or to simply say in the abstract that time was shorter with your tool, but refrain from this kinds of inaccurate statements.

Introduction

- Do you have a reference for this sentence: 'This can make the process very time consuming, even for small reviews; the labour required is obviously compounded in very large reviews, such as those seen in preclinical research.'
- Information provided later on, about 6000 studies for a corresponding research could be provided here in the Introduction, i.e. moved from its current position in the manuscript.
- -It would be beneficial to explain in the introduction what is the motive for this work when free and easy-to-use tools such as Plot Digitizer exist?
- Please elaborate further would be a motivation to make 'a tool which is designed specifically for data extraction in the context of systematic reviews'. It could be argued that it is irrelevant what is the purpose of data extraction [perhaps to enable some functionalities that the desktop software does not have?]. Perhaps, also to mention methodological research, i.e. research on research – there are potentially more types of studies that would need data from figures.

Research questions

- Please identify who are 'participants' here.

Identifying graph types

- 'we selected 34 papers identified in the context of systematic': it would be good to provide the list of those papers in the supplementary file

Design

- 'would be extracting data from a PDF file': I would remove emphasis on pdf files. Or clearly specify that this tool is only for figures in PDF files.

Evaluation design

- I would suggest to explain already here how many participants you had in this evaluation exercise

Discussion

- Sentence: 'Our study was not designed to show superiority, but suggests that there may be an average saving in time of around 6 minutes per graph': the same comment for this time as in the abstract. I would only write about time savings but not provide here minutes per graph. Not every graph will be the same, and then the savings will not be the same for each one.
- Can you explain why is the new tool more accurate than the electronic tools that the participants used as their current method? What is it in those other electronic methods that is inferior, compared to the new tool?
- Kadic et al., should be corrected to Jelcic Kadic et al.
- Section where the manuscript of Jelcic Kadic et al. is mentioned should be revised. The Jelcic Kadic et al. manuscript describes a study where electronic data extraction tool was compared to a paper-based method. However, here paper-based method was excluded from the comparator methods. Therefore, in this study the authors have compared a new electronic tool with existing electronic tools.
- I would revise 'graphical data extraction tool to current methods of data extraction' into: 'new data extraction tool, described hereby, to several desktop electronic methods of data extraction that are currently in use'.
- It would be nice if the authors could come up with a name for this new tool, so that we do not have to refer to it as a 'new graphical data extraction tool'.
- I would suggest revising all these 'nearly 6 minutes in time' from the Discussion, to simply talk about less time needed with the new tool.

Future work

- Integration with RevMan would also be nice.

Nice work, overall. Congratulations to the authors.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: I was the principal investigator of the Jelacic Kadic et al. manuscript, cited in this study.

Reviewer Expertise: evidence synthesis, research methodology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 02 Mar 2019

James Thomas, University College London, UK

We would like to thank Livia Puljak for her helpful review and suggestions for improving the paper. We have made the following changes.

Abstract.

1. *We agree that PDF is not the only format / medium used, but this is the area where we see most need, and made not change to the abstract. However, we have clarified the scope of our evaluation on page 1: "While we acknowledge that there are a range of different ways that papers are presented (from online to paper only), we focused this work on the challenge of extracting data from papers published using the PDF file format. This format is the most ubiquitous electronic publishing medium for papers and we therefore see greatest efficiencies from improving software support in this area."*
2. We have clarified that we're referring to participants' standard practice for extracting data from graphs.
3. We have clarified that 'standard practice' is referring only to extraction of data from pdf documents.
4. We have added the number of participants to the abstract.
5. We have replaced 'prior preferred current approaches' with 'standard practice' to be more

consistent.

6. *We have revised the final sentence to read: “Our study was not designed to show superiority, but suggests that, on average, participants saved around 6 minutes per graph using the new tool, accompanied by a substantial increase in accuracy.” We don’t think we can go as far as to say how many seconds / minutes were saved per data point, as there are other factors when considering how long data extraction takes.*

Introduction

1. I’m afraid we don’t have a reference for that sentence – it is based on personal experience and observation of very many other reviews!
2. We tried moving that text higher up, but it then spoilt the flow of the argument in the section where it was before, so have left it where it was.
3. “It would be beneficial to explain in the introduction what is the motive for this work when free and easy-to-use tools such as Plot Digitizer exist”. Thank you – that was indeed a missing piece of our argument! We have added a section into the penultimate paragraph of the introduction to address this point. (This addition also addressed the next point.)

Research questions

1. “Please identify who are ‘participants’ here.”. In the research questions, we already say that they are people *with a pre-existing interest in and knowledge of systematic review. We have now moved the “Participants and Recruitment” section directly under the “Research Questions” section so that the reader has this information earlier on.*

Identifying graph types

1. ‘we selected 34 papers identified in the context of systematic’: it would be good to provide the list of those papers in the supplementary file. We have added this information in an additional supplementary file.

Design

1. As clarified above – we have limited the scope of this evaluation to pdf files.

Evaluation design

1. We think that this information belongs in the results.

Discussion

1. We have revised in the same way as above.
2. “Can you explain why is the new tool more accurate than the electronic tools that the participants used as their current method? What is it in those other electronic methods that is inferior, compared to the new tool?” That is a great question – but we can’t unfortunately! We didn’t investigate this question in the evaluation.
3. We have revised to clarify that the Jelicic Kadic study compared paper-based to electronic media.
4. We have revised that sentence as suggested.
5. We agree that not giving the tool a name is a bit awkward in terms of writing the paper. This was a deliberate choice however, as the tool itself is proof of principle, and not something that systematic reviewers are able to use at the moment.
6. We can only see one other mention of ‘6 minutes’ in the discussion (apart from the one already revised) and think that this formulation is an accurate reflection of the results. Future work RevMan is really a platform for meta-analysis and review publication, and we do not think its

development roadmap includes data extraction (Cochrane uses other tools for this), so haven't suggested that such a tool should be integrated within RevMan.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 21 December 2018

<https://doi.org/10.21956/wellcomeopenres.16057.r34418>



Chris Marshall

Institute of Health and Society, Newcastle University, Newcastle, UK

Well written paper, but I'm confused as to where the novelty actually lies here. WebPlotDigitizer is an existing and already very useful tool, and it's not clear to me what the value of this 'adapted version' actually is? Is it the ability to view the PDFs with the graphs directly in the tool? Not sure. I think you need to make clearer the differences between the original software and this evolved version. Also, no proper citation to WebPlotDigitizer!

I'm also not sure as to what is actually being assessed here: is it an evaluation of this particular tool or are you assessing the concept of using a graph extraction tool to save time (or is it both?). If it's the former, it would be good to see a more direct comparison with other similar tools in this space.

Really appreciated all the links to supplementary files, raw data and code etc. Excellent in terms of transparency and reproducibility.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Health, evidence synthesis, systematic reviews, systematic review software/automation

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 14 Jan 2019

James Thomas, University College London, UK

We would like to thank Chris Marshall for his helpful feedback. We have modified the article and provided detailed commentary and screenshots which show the novelty of our approach in more detail (we agree that we should have included this in the original). We have also provided a more explicit statement of our research objectives.

Competing Interests: None

Referee Report 11 December 2018

<https://doi.org/10.21956/wellcomeopenres.16057.r34420>



Jon Brassey 

Trip Database Ltd., Newport, UK

An impressive paper with just minor points...:

- 1) "Systematic review and meta-analysis are research techniques whereby all available literature on a research question is collated and analysed to give an overview of that field." I'm thinking that it should include something like "...an attempt is made to locate and use all available literature".
- 2) "there is the clear potential to improve both the speed and accuracy of extraction of outcome data from the included studies once the required outcomes have been identified" - I'd say it's not clear if you're not immersed in automated methods.
- 3) Cochrane collaboration - hasn't it dropped 'collaboration'? I appreciate you made the 'c' in 'collaboration' lower case!
- 4) "...and are used frequently in clinical decision making" - can we have a reference please?
- 5) Because of the challenges in manual data extraction from ROC/AUC graphs, these were not offered in this set - I'd have this as a new paragraph.
- 6) Exploration of drop-outs. 32 signed consent and ten completed the trial - I'm sure it's for mundane reasons - but might this high-level of drop-out be down to some unforeseen reason that might impact future roll out of the technology?
- 7) Possibly not one for the paper but I do think the long-term future is ensuring the trial variables (including

data used in plotting graphs) should be embedded, in computer-readable format, in the articles meta-data. That removes the problem.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical search, automation, question answering, evidence synthesis

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 14 Jan 2019

James Thomas, University College London, UK

We would like to thank Jon Brassey for his helpful feedback. We agree with all his comments and have modified the article to integrate them all - with the exception of the last one: we agree, but this requires longer-term changes to publication practice!

Competing Interests: None

Comments on this article

Version 1

Author Response 20 Dec 2018

James Thomas, University College London, UK

Hi Ankit,

Thanks for your comment. It is unfortunate that there's a bit of a delay in moderation of these comments, as my guess is that you couldn't see my response to Geoffrey before commenting yourself.

I do remember checking for a proper citation for WPD, so apologies it wasn't in the paper - but will be in the next version.

When we did this work (back in 2016), we integrated pdf.js and WPD in what we hoped would become a useful tool, where you could upload a pdf, draw a box around the graph, and then extract the numeric information. As we integrated two tools, and did a lot of additional coding, I think it reasonable to describe the result as a new tool - given that we are clear where the main components came from. I think technology developments have overtaken us a bit, so a future tool to support systematic reviews may use, as you say, just the pdf support that's now in WPD and not need the pdf.js component too. One of the main eye-openers for us when doing this study is just how difficult - and unreliable - our current approaches to extracting data from graphs are, so we fully appreciate and recognise the important contribution to fixing this that WPD makes.

Best wishes, James.

Competing Interests: Author on the above article

Author Response 14 Dec 2018

James Thomas, University College London, UK

Thanks Geoffrey, we will fix the lack of citation when we revise the manuscript. I remember looking for information re how to cite WebPlotdigitizer, and am not sure why it didn't make it into the submitted paper.

Competing Interests: One of the paper authors
