We thank the editor and reviewers for their feedback and the opportunity to revise our paper.

Overall, we have prepared a revision with changes primarily in the analysis and discussion sections (we point out specific changes, with page numbers, in responses below). The major changes are:

* We have added statistical comparisons throughout, especially confidence intervals for estimating the population levels from our sample results,
* We break out proprietary and other kinds of software, examining how different kinds of software are mentioned.
* We break out functions of citation by the different kinds of software mentions, showing which functions different kinds of mentions are good at and which they show problems on.
* The flow of the analysis is now different: we first establish our classifications of software mentions, then our classifications of software. Then we turn to examining functions of citation. We think this improves the readability of the analysis.
* We have carried changes in our analysis through to the discussion and now include additional suggestions (e.g., projects ought to provide requested mention forms, journals and other styles should provide fallback mention forms).

Below we respond to each reviewer comment, highlighting changes that we have made in response.

Referees' Comments to Author:

Reviewer: 1

Comments to the Author

This is a well-conceived and executed work about a the problem of uneven citation of software used in reporting scientific research. The authors explain and justify antecedent research, theory, methods, and coding schemes clearly (and convincingly). The problem of software citation as described by the author is novel, challenging, and of significance to the broader scholarly community. The subject is important; the questions & thoughts I list below arise from my view that this preliminary work needs to be extended and that the paper needs to more forcefully make its point.

We thank the reviewer for their supportive comments.

The division of articles into "strata" or "tiers" based on impact factor appears to be logarithmic. Could this influence the analysis?

We have revised our discussion of impact factor selection to make our intent more clear, see page 6/7. We are seeking to provide an overall view of the state of practice in the literature, while still acknowledging that some of the literature is, in an unfortunately crude sense, more important than others, particularly in terms of likelihood to influence how others mention software (e.g., it is more widely read). Of all the tools at our disposal impact factor seemed the least problematic. We choose a sampling technique weighted towards higher impact factor journals because we think that if we were to describe practices without doing so, a fair enough response would be to say, “Yes, but your sample included lots of lesser journals; what really matters in terms of what people listen to and read are the higher impact journals”. On the other hand, it is interesting to see what is happening in less impactful journals (especially as there is some reason to believe that higher impact journals police their citations more closely, including for number of citations). We aimed to strike a balance and think our method does a defendable job in doing so; we hope others agree.

There are instances in the text (pg 6, pg 17) in which the authors equate "high impact factor" and "high quality". I am not certain that the two phrases are synonymous and would urge caution in the use of such value-judgment laden verbiage.

We agree and are sorry that this language crept into the submission; those value judgments are not relevant to the paper and we have removed them. We are happy that the reviewers pointed this out.

The authors selected 3 sets of 30 articles from each of the impact factor tiers. Of these, a smaller number were appropriate for subsequent analysis. This is (these are) a small sample(s) - is it still significant? (Certainly, intra-strata comparison would be challenging with such small samples.)

This is an excellent question and an oversight on our behalf. We have worked through the paper and now show 95% confidence intervals showing the likely range of occurrence in the population of biology articles, given the occurrence in our sample. For the overall sample these are reasonable (e.g., plus or minus 5%), and, we think, give useful insight into the proportions of mentions in the population that have particular forms, include particular details or facilitate finding and using software. In other cases, at the level of the full sample, we do not think it is large enough and do not interpret our results that way (for example, we do not provide a “league table” of software used).

On the other hand, for intra-strata comparisons (being a 3rd of the size) our confidence intervals are quite large. We never intended our stratification to support comparisons between strata, indeed we don’t make hypotheses about this. The stratification was to produce balance for the overall sample. Yet we had fallen into making comparisons between strata. In the revised submission we de-emphasize these, but continue to show results by strata primarily to help the reader assess the quality of our effort at producing a balanced sample. In each case we have now provided confidence intervals that show few cases of statistically significant differences across strata (e.g., the declining use of the “like instrument” citation style in lower impact factor strata). Yet, we do not rely on results finding no differences for our conclusions; it may be that our sample is simply not big enough. Overall we think that this work improves the paper, ensuring that we are as straightforward as possible about what we are claiming and what we are relying on in our conclusions.

In the end we think the sample is sufficient to draw attention to the issues found and appropriate for the first study in this area. Most importantly, we do think that the sample is large enough to advance the policy question that is at the heart of the paper: software mentions are very diverse and underspecified, leading to a diversity of practice and clear evidence of problems.

We have adjusted our discussion of sampling to make our logic clearer, see page 6.

If these results are to be extended past the field of biology, the sample may need to be larger, or the sample significance verified. Also, given the great difference in performance among the tiers, I wonder whether the journals among the tiers can be compared to each other without considering other characteristics such as institutional affiliation or the nature of the research reported. (i.e. are software-heavy projects disproportionately coming from better funded/higher profile sources? Do they come from sources that have greater access to computational facilities/personnel?) It would be valuable to know if these results are consistent across disciplines. The problem of software citation is relevant across scholarly communication - is the problem similar in areas outside of the sample?

We think that these questions are interesting and appropriate for follow-up research; as discussed above our goal was not to make statistically valid comparisons between tiers but to ensure balanced coverage while describing the diversity of how software is mentioned. We would love to pursue these questions in future research (or to assist others in doing so, including by sharing our data, coding scheme, and analysis code).

The authors used a very sophisticated data management system, so their reliance of relatively basic statistics is a surprise. If it is deliberate, the reader should be informed.

Again, we thank the reviewer for pointing this out; we think the addition of confidence intervals improves the paper and hope the reviewers agree.

The results presented are largely descriptive (incidence of citation, nature of citation). In the spirit of extensibility, I would love to see a degree of comparison between disciplines, (perhaps comparing top and second-tier journals in similar, but distinct areas, such as bio informatics and health informatics.) Are there correlations or hypothesis testing to be done? Can these results be extended across the "Republic of Science"?

We would love to know as well! Unfortunately we can’t do everything in this one paper and remain timely and relevant to the fast-moving changes in scholarly communication. We hope that our study will prompt and support further work, including comparisons across disciplines. We especially hope that by making our reliable coding scheme (and analysis software) available to others we can make future comparative work easier.

This paper is significant and well-written. As I read, I kept wanting more - more sample size, more statistical analysis, and more comparison. By expanding the scope of the comparison, I believe that the paper could compel interest more broadly - and the subject is certainly of sufficient importance to warrant such treatment.

In this version we have added a more intuitive classification of software types and an examination of how different kinds of software are mentioned. Probably more interestingly, we examine how different means of citation perform in accomplishing different functions. We example we can now ascribe the absence of crediting information solely to informal citation forms and highlight the almost complete absence of version information when the “Cite to publication” form is used. We think these additions are useful and hope the reviewers do too.

In summary, I feel that this research needs to be made more forceful. As reported, the results address a small subset of the biomedical literature. These limited findings hint at a serious issue in scholarly communication - I would like to see this issue addressed in such a way that the importance to the scientific community in general were easy to see.

We appreciate the reviewer’s interest in the material and do hope to expand this work in the future, beyond the biology literature and with larger sample sizes. Yet we think that the current scope of the study is appropriate for the first article to address the topic and sufficient to support its claims. With this as a starting point, and the coding scheme and the questions for future research, we hope to attract other researchers to drive forward our understanding.

Reviewer: 2

Comments to the Author

The manuscript by Howison and Bullard (How is software visible in the scientific literature?; JAIST-2014-08-0571) presents and analysis of software mentions in scientific literature. While the manuscript undertakes an analysis of an important, but relatively neglected part of the scientific literature, I have several concerns about the work as presented.

We thank the reviewer for their comments and their concerns below.

1. The authors group ‘software” as a single entity, and include both commercial (for profit software) and freely available and open source software. Authors note that three main systems of citation are in place (“cite to publication”, “like an Instrument” and “name only”), but there seems to be no distinct analysis of whether commercial software is generally cited like an instrument, while open source is cited like a paper or by name. These would seem to be the appropriate methods of citation. Indeed, a question the authors should address is whether commercial and open source/freely available software can or should be cited in the same way?

We thank the reviewer for this suggestion, on reflection it is a natural comparison to make. First we needed to separate and make more intuitive our classification of types of software, now shown in Fig. 6 on p. 20. We can then undertake an analysis of how different kinds of software tend to be mentioned, shown in Fig. 7 on p 21. Indeed commercial software is more likely to be cited “like instrument,” than other kinds of software are. Yet, somewhat surprisingly, it is not the case that most Proprietary software is referred to using the “like instrument” category: informal mentions remain the most used category. Overall there remains a wide diversity of mention styles. This strengthens the case we now make later in the paper for software to be clear how it ought to be cited, and for journals to have “fall back” policies. We thank the reviewer for the suggestion and hope our additional analysis responds usefully.

In terms of appropriateness of the “like instrument” category, we think there are two questions here. We agree that the higher prevalence of “like instrument” does show that authors tend to attempt to use the preferred form of citation, although clearly do not in many cases. The secondary question is whether the “like instrument” form accomplishes the functions of citation more than other forms. We now examine this later in the paper. We find that the form is good at including some crediting information (usually to firms, not individuals) but no better than others at functions like including version numbers of software. We hope our analysis is clearer at exploring these interesting questions.

2. While I can see that “accessibility” of the software (i.e. being able to find the software) is a component of “visibility” (and replication), I disagree with the contention that the ability to modify source code is an aspect of “visibility”. In this regard, the authors blur the lines between “visibility” and “utility”. Again, lumping commercial and non-commercial software together seems overly simplistic.

The distinction between visibility and utility is an excellent point and we thank the reviewer for this comment. It is analogous to the distinction between studying citation of and the usability of scientific literature: both are important to the overall conduct of science, but “visibility” doesn’t cover both. We have revised the title and the language in the paper to make this clear: we are studying visibility but also (albeit in a more limited way) the utility of software that a reader might obtain.

We agree that simply because something is only available for payment doesn’t mean that it is not useful: that’s why we break out accessibility into multiple categories. We now more clearly differentiate between open source and non-open source software (in discussing accessibility and modifiability). We now also cite authors and studies that provide evidence from scientists arguing that open source has higher usefulness for replicable science (e.g., Howison & Herbsleb, 2011; Ince, Hatton, & Graham-Cumming, 2012; Stodden et al., 2010).

3. The authors note that in their selection of 90 articles, 59 articles mentioned software, but 31 did not. However, the authors did not apparently further explore whether the articles that did not mention software should have cited software. Was there evidence in these papers that software should have been mentioned? (Examples would include evidence of statistical analysis or complex data generation).

We agree that this is a very interesting question; we (and many software package authors!) would love to know when software has been used but not mentioned at all. While there are some likely “giveaways” (as the reviewer mentions, graphics and statistics) we could not come up with a reliable method of assessing the general question (which applies in all the articles, not only those that don’t mention any software). We therefore left “non-mentions” out but are exploring that in further work. We highlighted our decision in the paper on page 3 and 9 and call for additional work in this area in our conclusions.

4. While the authors explain their selection of three strata for journals, I am not sure that this is really justified in the context of the study (“visibility in the literature”). Firstly, in terms of publishing weight, 90-99% of all publications are published in the authors tier 3 journals. What is going on in effectively 5 (tier 1) journals really does tell much about the wider field of science. In this regards papers in Tier 1 journals tend to have extensive editorial input, and so again are less reflective of practices in the wider field.

Our selection of three Tiers is to balance broad coverage of the wider field with sufficient coverage of those venues most looked at and discussed. Due to the concentration of readership and reputation in scientific publishing we’re essentially trying to sample from a power-law or at least long-tail distribution, a difficult task. That’s why we turned to stratification, trying to balance interest in the higher tier journals (“important science”) against wider coverage of the whole field (including specialty journals). It’s a balancing act, but we think it’s better than having readers dismiss the findings because they were either only the “top journals” or they were from “journals no one reads” or “very specific sub-fields”.

Incidentally, the fact that the articles in Tier 1 came from only 5 of the top 10 journals was somewhat disconcerting, but was the result of the randomization, so we thought it best to leave it alone.

In this regards, authors also need to choose their words more carefully. While there is no objection to the use of “higher impact factor” (or indeed “lower impact factor”, the use of “higher and lower quality” (e.g. page 6, line 55) and “higher quality” and “lower quality” (Page 17, lines 53 and 54) should be avoided.

Thanks for pointing this out. Indeed we definitely did not mean that and are annoyed that those errors made it through to the text. We have changed these instances and looked hard for any others. We have also made sure to always use “journal impact factor,” even when discussing articles.

Indeed, given the considerable reservations towards impact factors (with impact factors generally being the product of a relatively small proportion of a journals total paper output), the use of impact factor for stratification would appear to be relatively outdated. A more representative picture would have been obtained by selecting equal article numbers based upon stratification by journal quartiles, or, ideally through the use of article specific metrics.

We acknowledge the issues with Impact Factor across the board, especially when used for assessing quality of individual articles and scholars. In this case, though, since journal policies are relevant to how software is mentioned we felt that the journal unit of analysis was appropriate, even useful. Article specific metrics are not available across broad areas of the literature (hopefully that will change in future!) Journal quartiles would have been another appropriate method of stratifying (also based on journal impact factor), but we hope the three tiers are easy enough for the reader to understand.

5. The issue of “improving credit” (page 20) again would seem to link poorly to the central theme of the paper “visibility”. While there is certainly a case to be made that credit should be appropriately assigned, is this an issue of “visibility”? Again, should commercial and non-commercial software be held to the same criteria?

One can see visibility from two angles: from the perspective of a reader seeking to replicate or extend the science reported in a paper or from the perspective of the authors of the software. In the first case, yes, credit is secondary to finding and using the software, but in the second case (for developers of the software) being credited is crucial to their careers and so is definitely related to visibility.

Definitely credit of this kind is more important to academic authors than to commercial providers. We mention this now on page 25 (while pointing out that just because software is sold it isn’t the case that authors aren’t interested in contributing to science; Microsoft, for example, has an entire division devoted to scientific use and sponsors conferences like e-Science). We also raise the question of whether it’s sensible to have separate standards for commercial software and software written in the hope of scientific credit and reputation. It’s certainly hard for an end-user to know, so projects most likely should be explicit about how they’d like to be referenced. We now discuss this on page 26.

6. The authors reference a number of pieces of software used to create graphs or to analyze data (Acknowledgements). I was unable to find a single version number, and the authors themselves use a number of citation styles. This would have seemed to be the ideal place for the authors to reference the software they used in the manner in which they advocate. Indeed, some clear examples of what the authors themselves think of as “ideal” citation styles would have been extremely helpful.

Indeed, thanks for pointing that out; see even those attempting to do the right thing are unsure! We have added these details to the Acknowledgements (moving them from the linked repository that provides all of our data and analysis code). We are hesitant to propose “ideal” solutions at the moment, believing that some “user testing” would be needed. We hope to work with the other researchers mentioned on advancing that in future.

7. No test of statistical significance has been applied to any of the data, and this should be corrected. Are the differences seen between strata statistically significant or not significant?

On reflection we agree; we have added statistical information throughout, primarily confidence intervals showing appropriate estimates for the proportions to be found in the population of articles. We appropriate we have also used chi-squared statistics to show relationships (such as between types of software and types of mentions, and between types of mentions and functions of citation). We included inter-strata comparisons primarily to show the reader that our effort to balance our sample had succeeded; however we now include confidence intervals and point out a few places where there are clear statistically significant differences across strata (such as the “like instrument” form being used less in the lower impact factor journal tier). Overall, though, we do not make hypotheses or draw conclusions about inter-strata differences.

In the longer term we think that assessing correlation with journal impact factor is not a particularly interesting goal; and, honestly, we think that it would be more interesting to assess correlation with likely causational factors such as specific journal policies and/or author’s beliefs about the role of software in science. We make those suggestions for future research in the discussion and conclusion.

We thank the reviewers for pushing us to improve our analysis; we hope they agree that the paper is more informative now.

8. Titles of papers should generally be “informative”. In this regard the title of the manuscript is not informative - it tells us what the study is about, but not what the study found. The title should be re-written, avoiding structuring the title as a question.

We re-wrote the title:

*Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature.*

We would appreciate reviewers or editors suggestions for improvement.

9. References need to be re-checked. See reference to “Loo, M.P.J van der” for an example of incomplete referencing.

Our apologies; we have checked and re-checked the references and believe they are now all in order.

Overall, we think the reviewer’s feedback was very useful in improving the paper and hope that the reviewers and editors agree.

Sincerely,

James Howison and Julia Bullard, UT Austin.

**References**

Howison, J., & Herbsleb, J. D. (2011). Scientific software production: incentives and collaboration. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (pp. 513–522). Hangzhou, China. doi:10.1145/1958824.1958904

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488. doi:10.1038/nature10836

Stodden, V., Donoho, D., Fomel, S., Friedlander, M., Gerstein, M., LeVeque, R., … Wiggins, C. (2010). Reproducible Research. *Computing in Science and Engineering*, *12*(5), 8–13.