The limited value of journal peer review in public health: a case series of tobacco harm reduction articles

Working Paper version, 23 June 2015

Carl V Phillips, Chief Scientific Officer, The Consumer Advocates for Smoke-free Alternatives Association (CASAA)

Igor Burstyn, Associate Professor, Drexel University School of Public Health, and Member of CASAA Board of Directors

Brian L Carter, Executive Consultant, Carter and Day LLC

Abstract

Background: A widespread belief holds that the journal peer-review process has magical powers to ensure that published claims are correct. While this misperception has limited consequences in many fields, in public health it results in consumer, clinical, and policy decisions being based on blind faith in the accuracy of published claims. At best, the review process is merely a couple of readers -- perhaps, but not necessarily, highly expert -- reading through a paper to ensure the research and presentation are reasonably sound. In reality, even this is often not accomplished.

Methods: We conducted reviews of 12 articles that focused on tobacco harm reduction published in a mainstream public health journal, *BMC Public Health*, consecutively during 2012-15. We each wrote a reviewer report of the manuscript version that was sent to the journal reviewers, as if we were writing a review for a journal. We then compared these to the reviews written by the journal reviewers. Additionally, we reviewed the changes made to the papers as a result of the journal reviews.

Results: Almost all the papers in the dataset suffered from major flaws, most of which could have been corrected, but none were corrected by the journal review process. The journal peer reviews were almost all inadequate and many contained no substantive comments. Those that contained substantive observations still did not identify most of the blatant major flaws that we noted. In the single case where a journal reviewer identified many of the major flaws, the comments were basically ignored by the authors and the paper was published with no substantive changes. Other than cosmetic improvements, the journal review process was about as likely to make the published version worse than the submitted manuscript, rather than better. Papers with no apparent value were published by the journal and the potential value of other studies was lost

because serious flaws in the paper were ignored. Unreported conflict of interest was common among both authors and reviewers.
Conclusions: Faith in the journal peer-review process is misplaced. Even at best, the process cannot promise that a published claim is correct, but in reality it does not even ensure that patent major flaws are not present. In public health, the phrase "according to a peer-reviewed journal article" seems to mean little more than "I read this somewhere."

Background

Many who are unfamiliar with the journal peer-review process tend to fetishize it as having magical powers to ensure published research is properly designed and expertly analyzed, and that the resulting conclusions are trustworthy. As with acts of magic in stories, these consumers of journal articles have no model for how peer review accomplishes that, but they accept the fiction that it just works. Those who have personal experience with the process know what really occurs, and thus should recognize the limitations, but many of them similarly overstate its value.

The journal peer-review system was a relatively recent addition in the history of science, from the mid-20th century. It was useful during the period when science had broadened and compartmentalized beyond the ability of most readers and editors to be experts in all relevant details of a science, but when a few individuals with such expertise about a particular paper could be reliably recruited to review it. But due to increasing productivity, knowledge, and specialization, that time has passed in many fields. Thus, in many sophisticated sciences the process is being supplanted by information-age alternatives that facilitate broader review, collecting feedback from many experts after making the paper available to all interested readers (e.g., arXiv.org). But in others fields, the limited journal review process is still dominant and increasingly strained.

Misunderstanding the limits of journal peer review is of limited importance for readers dabbling in fields that have little or no worldly consequence (e.g., ancient history, ethnomusicology). Believing in the magic of the journal review process is unlikely to occur in fields where most serious consumers of the content have enough expertise to judge the material for themselves and also read the wider literature that provides analysis that may be absent from journals (e.g., agricultural science). But in the health sciences, many published articles – or at least their abstracts and stated conclusions – are widely read and accepted by people with limited expertise who uncritically base consumer, clinical, legal, and policy decisions on them because of their peer-reviewed status, with no additional scrutiny of the quality of papers or legitimacy of the conclusions. The health sciences also lack quality control measures that exist in other fields. For example, in economics, another field where research has worldly consequences, few important

papers are submitted to journals until they have been circulated as working papers and the research has been presented in multiple seminars, and the resulting assessments of many expert reviewers have been incorporated into rewrites. In the health sciences, by contrast, articles are often published with only a handful of people beyond the authors ever having seen the paper or even a presentation about the research.

Journal reviewers typically have access to only the same paper that is available to the journal's readers. It is almost impossible for any reader of a manuscript to detect the most egregious problems that are found in journal publications, such as fabricated data or errors in data collection and coding. More subtle but equally serious problems, such as unreported multiple hypothesis testing [e.g. http://jech.bmj.com/content/63/8/593.short; http://www.ncbi.nlm.nih.gov/pubmed/15296515], will only be recognized by the most skilled reviewers and even then are difficult to confirm unless the reviewer asks for additional information and the editor and authors comply. The potential for undetectable errors, accidental or intentional, is greater in public health than in many other fields of research [https://www.researchgate.net/publication/267304455_Peer_Review_in_Epidemiology_Cannot_Accomplish_Its_Ostensible_Goals_Due_to_Incomplete_Reporting_and_Unverifiable_Analyses. And, of course, a researcher might do everything right and report it fully, but a particular study result might still misrepresent the real-world phenomenon it is attempting to measure.

With these inherent limitations in mind, we must hope that at least the editor and couple of reviewers who see a submitted manuscript – often the only outside readers to ever review health science research before publication – provide what critical evaluation is possible, and flaws they identify are corrected. All we can realistically ask is that the journal review process ensures that the paper is *prima facie* correct and ready for assessment by the community of experts. Note that this observation in itself represents roughly the opposite of the belief of those who idolize the process, who act as if it represents the final word that a paper is accurate and valid, not merely a rough screen that it is probably not glaringly flawed.

Journal reviewers can easily identify some types of methodological flaws, failures to adequately report methods and results, some types of errors in the reporting of results, conclusory claims that do not follow from the results, claims of implications that do not follow from the research, and if sufficiently expert in the topic matter, background claims and premises that are dubious or inaccurate. However, because flaws of these types frequently appear in public health journal articles, there is some question about what actually occurs in the journal review process.

The motivational example for the present research is an article by Popova and Ling that appeared in *BMC Public Health* (BMCPH) in 2014 [http://www.biomedcentral.com/1471-2458/14/997]. The authors conducted a study that had no chance of generating any useful information and then

stated aggressive conclusions and policy recommendations that were in no way supported by the research. There were serious concerns about the ethics of their research. These were described at length by CVP and Clive Bates (CB), and some related analysis was written by BLC [http://antithrlies.com/2014/09/26/new-public-health-research-lying-to-people-can-affect-them-as-if-they-didnt-already-know/ http://antithrlies.com/2015/05/12/the-failures-of-peer-review-do-not-begin-with-the-journal-more-on-the-popova-ling-fiasco/ http://www.clivebates.com/?p=2418 http://www.biomedcentral.com/1471-2458/14/997/comments (latter should probably be three referenes, one to each comment]. In particular, CVP wrote what a peer reviewer might be expected to provide to a journal and contrasted it with the contents of the journal reviews [http://antithrlies.com/2014/09/27/what-is-peer-review-really-part-4/]. The journal review of this paper, a paper which was composed almost entirely of obvious flaws, failed to correct any of the problems and actually made the final version worse than the submission.

One might be inclined to assume this was an outlier. But given the lack of systematic analysis of the journal review process in public health, it is difficult to be sure. Sometimes glaring failures are extreme rarities, but sometimes they represent the most easily noticed examples of a common phenomenon.

To assess this point and to further educate readers about what constitutes typical journal peer review in public health, we conducted our own reviews of original manuscripts submitted to a journal, compared our reviews with the actual journal reviews, and assessed the effectiveness of the journal review process in improving the final published articles. The primary purpose of this research is to explore the true implications of a paper having “peer reviewed journal article” statue and to begin to document the extent to which the process fulfills even its limited potential in public health publishing.

Methods

We chose to analyze articles about or related to tobacco harm reduction (THR), the substitution of the use of low-risk alternatives (smokeless tobacco, e-cigarettes, pharmaceutical nicotine products (NRT)) for smoking. This choice was based on it being our common area of substantive expertise, its being the subject of the motivational example, and the variety of types of studies that are done in this subfield. We defined an article as meeting this criterion if the abstract mentioned THR (with or without using that phrase), or if the abstract mentioned smokeless tobacco or e-cigarette products, since their major role in the world is THR. Since NRT is primarily studied as an aid to becoming abstinent rather than as a harm reduction alternative, mentions of NRT were not sufficient for inclusion, but articles specifically about using NRT for THR were included.

We chose to study a BioMed Central (BMC) journal because, to their great credit, they publish the original submissions, reviews, and revisions alongside the final articles. This is a major improvement over the anonymous and hidden peer review process used by most health science journals, and by itself should tend to encourage better reviews. However, this better methodology cannot achieve its full potential without some effort to audit what has been done, as we contribute here. In addition, being online-only journals, the BMC journals do not suffer from the problem of print journals wanting to limit the length of articles so they can physically fit more of them (the typical print word limits are perhaps barely sufficient to report on a simple study, like a clinical trial, but are grossly inadequate for reporting more complex research). Reviewers for print journals may excuse author failures to adequately report as being unavoidable due to the word limit, but this should not occur with a BMC journal. We limited the review to BMCPH because it published the motivational article, for simplicity, and because it is an obvious target among BMC journals for papers on our chosen topic matter. It can be considered a neutral publisher of THR articles, with a history of publishing articles that are considered among the most important in the field as well as some that are actively hostile to THR; this contrasts with such BMC journals as *Tobacco Induced Diseases* and *Harm Reduction Journal*, which would be expected to attract only articles which focus on the negative or positive views of the topic matter.

We chose to include the 10 most recent articles in BMCPH, in addition Popova and Ling, that met the selection criteria as of 3 March 2015 and to continue to monitor the journal for new qualifying publications that appeared until we completed our reviews. The quantity was chosen to strike a balance between coverage and workload; since this research approach does not facilitate quantitative claims, there was no issue of "power" or such. We expected, based on experience, that our dataset would provide examples of a variety of methodological approaches and be sufficient to provide enough examples of various common limitations of peer review. It is a systematic sample, though necessarily not representative. Indeed, it is not even clear how one would define a target population to represent.

The search was conducted by one of us (CVP) doing a rough screen – searching all BMCPH articles that contained the word "tobacco" or "nicotine" – and then manually inspecting abstracts in reverse chronological order until enough were found. We were forced to make an *ad hoc* adjustment to the protocol when we discovered that one qualifying article [http://www.biomedcentral.com/1471-2458/13/210] had been published without any external reviews; we eliminated it from consideration and continued to next most recent article, for a total of 10 [REFS appear in Table 1]. One additional qualifying article was published during our review process [http://www.biomedcentral.com/1471-2458/15/244]. These, along with Popova and Ling, were analyzed. The list of articles appears in Table 1 (https://docs.google.com/spreadsheets/d/1Mq-bucQhgeTAK2cWlWZw7Fiy1Xt0xNbm1FISLeuAo_E/pubhtml), and hereafter they are referred to by the name of the first author for

convenience. The papers were submitted between March 2012 and July 2014 and published between November 2012 and March 2015.

We wrote reviews of each submitted manuscript as if invited by the journal to review it. This was done by each independently, and without having seen the journal reviews. Each manuscript was reviewed by CVP, IB, and BLC, with two exceptions: Since much had been written about Popova already, we used the extant comments by CVP (who wrote a review that approximately followed the protocol of this study) and CB. One of the qualifying papers was written by IB and the submitted version had been revised based on extensive comments by CVP on a working paper version, and thus neither of us could offer further useful review. We recruited CB – chosen because he had already contributed a review of Popova – to join BLC in reviewing that paper. All the reviewers have general expertise in public health research and methods, and particular expertise on THR and THR products. We collectively brought extensive experience as highly respected authors in the field, as reviewers for journals, as journal associate and chief editors, and as teachers of critical research methods, and we have conducted formal research and investigations on the peer-review process. The reviewers also brought extensive expertise about important methodological points, including data analysis and statistical method (CVP, IB), qualitative research methods (BLC), toxicology and environmental chemistry (IB), social and behavioral science research (CVP, BLC), policy analysis (CVP, CB), scientific epistemology (CVP, IB), human subjects ethics (BLC), clinical and laboratory tobacco use research (BLC), and critical review and meta-analysis (CVP, BLC, IB).

The manuscript version we reviewed was the first one that was sent to external reviewers (in many cases a journal editor asked for minor technical changes to an earlier version of the manuscript before sending it out). Each reviewer was instructed to write the review they would for a journal with the following exceptions: ignore copy edits and similar minor problems that might be noted in a review; there was no need to explain subject matter or methodological points, as you might feel inclined to do for possibly inexpert authors or editors, as long as the basis of the point was clear; if reviewing that paper for a journal would consist of citing one or more fatal flaws and noting that there was no point in continuing, or the reviewer would have insisted on some particular revision before continuing to review, continue to provide further analysis of the content for the sake of our study aims. Further details were left to the individual reviewer's discretion, as is standard for journal reviews.

None of us read the others' reviews, discussed the reviews with one another or anyone else, nor read the journal reviewers' reports or any other analysis of the articles until our own reports were finalized. Collectively we had some familiarity with most of the articles in the dataset, but had never conducted or read a critical analysis of any other than Burstyn and Popova. It is notable the extent to which the enforced isolation was a frustrating restriction: Journal reviewers are

typically instructed to not show manuscripts to others, but serious reviewers will often ignore this and seek input from those with other expertise and often attempt to improve the value of their review by finding already published analyses of the work (though, as noted, in public health the latter typically cannot be done since the work may never have been reviewed by anyone before being submitted). In quite a few cases we noticed that we would have sought the input of others with particular expertise rather than working in isolation, though the latter is typical reviewer behavior so this made our process more similar to the typical process. We did not update or correct our reviews after our extensive further contemplation and reading one another's contributions. The frustration of not being able to improve the content of reviews following "I can't believe I overlooked that!" moments is itself an indictment of the process; there are inevitably points that a reviewer would have made had he thought of them, but contemplating a paper for less than a day, in isolation, is a recipe for overlooking many points that one would immediately agree with upon seeing someone else note them.

We then analyzed each of our reviews and the reviews by the journal reviewers and summarized what we believed were the most substantial points from our reviews and all the substantial points from the journal reviews. This summary appears in Table 2 (https://docs.google.com/spreadsheets/d/1ISZybAVVGiLEPw26AhEWnHEy5tMQxa8HEtkhjLN0apg/pubhtml). In the one case where there was a second round of substantive reviews (Edvardsson) we analyzed those also (we ignored follow-up reviewer reports that had trivial content). We used the authors' responses to reviewers and the final manuscript to assess what improvements were made to the paper.

The summaries in Table 2 are sometimes paraphrases of a specific statement that appeared in one or more reviews and sometimes were a summary of multiple specific points (e.g., "methods inadequately explained"). Details can be found the reviews themselves, which appear in the appendices that are linked in Table 1. Table 2 notes who among the reviewers offered a particular observation. The exact statement in a review is inevitably idiosyncratic, so we attempted to create summary points that consolidated functionally similar comments from multiple reviewers into one characterization. The lists in Table 2 include enough of the points we made to capture the essence of our reviews. (For papers with minimal serious problems, there is really not "essence" so much as details, and so we chose some highlights). We included *all* substantive comments from the journal reviewers in Table 2 to avoid understating their contributions. We omitted from the table any relatively minor points from one of our reviews that, upon discussion, we agreed were not a valid (they remain in the original reviewer report, of course). For the only reviews on which we had notable disagreement about major points (Shepperd, Edvardsson) our competing views are included in Table 2 and addressed in the text.

We compared our reviews to those of the journal's reviewers, with particular attention to whether we identified serious flaws they missed. As a secondary point, we noted the extent to which particular changes were made before the article was published. Assessing the quality and accuracy of the submissions and final publications themselves was not a defined aim of this study. However, it had to be done *en passant* in order to assess the reviews and review process, so a secondary result of this research is to provide such assessments, though those assessments are not systematic.

Results and Discussion

Table 1 briefly characterizes the studied papers and provides links to the published article (which contains links to the journal reviews and revision history via the "Pre-publication history" button). For readers' convenience, we combined the original submission, our reviews, the journal's reviews, and the authors' responses in appendices which are also linked from the table. Readers who are interested in more than the summaries of the reviews that appear in Table 2 and the text below can thus easily access the full content. The ordering of the tables follows the ordering of the analysis below, which was chosen to provide a useful narrative.

The level of detail of entries in Table 2 reflects the general quality of the paper because if we noted multiple fundamental problems, they crowded out detailed suggestions from the table (unless they were also noted by the journal reviewers, because due to the paucity of substantive comments from them we tried give them credit for any substantive comment). The detailed comments can still be found in the appendices for the fundamentally flawed papers to the extent that we noted them.

<u>Overview observations</u>

A mere glance at Table 2 shows how paltry the content of the journal reviews was. For 9 of the 12 papers, the most robust of the journal reviews contained only a tiny fraction of the substantive content of any of our reviews. (The exceptions were: Burstyn, where all of the reviews offered only suggested additions and clarifications at a detail level; Shepperd, where one of our reviews expressed serious concerns but two were more similar to the journal review, focusing only on details; and Silla, where one of the journal reviewers identified about half of the serious issues that we did.) We expect that those with faith in the journal review process assume that journal reviews have the level of thoroughness of our reports, rather than the level actually observed.

Moreover, for all papers other than Burstyn (where no reviewer noted any serious flaws) and Shepperd (where we disagreed as to whether there were serious flaws), our reviews concurred that there were flaws that were both patent and serious that were overlooked by all the journal reviewers. For all but two papers, we were in almost complete agreement about what the major flaws were (the exceptions are Edvardsson, where one of us noted positive aspects, but two of us

thought it was a complete disaster, and Shepperd where two of us thought it was largely sound but one had serious concerns). For some of the papers (e.g., Zyoud, Luo, Popova), we concluded that the research produced no information of value and the paper itself had *negative* value due to unsubstantiated claims. For others (e.g., Polosa, Silla) our reviews noted that the reported results clearly had value, but there were important flaws in the analysis and reporting.

The most common themes from our reviews were that the conclusions did not follow from the research and that the reporting of methods was grossly inadequate. Clear and major failures in each of those areas were evident in well over half the papers.

The methods section of a paper should strive to make clear to readers every relevant choice and action by the researchers. Someone who wanted to replicate the research, as closely as possible, should be able to do it based on what appears in the paper. More practically, since public health research is almost never replicated, a reader who thinks, "I wonder if they did X or Y, since it makes a difference in interpreting results" should find the answer. If these conditions are not met, reviewers should call for changes. Indeed, journal reviews almost always include calls for clarification of some specific details about the methodology, and our dataset was no exception. But we found that the reviewers seldom generalized their observation about inadequate reporting, calling for more adequate reporting of all methods; when they did it was ignored by the authors and this was allowed by the editors.

It should go without saying that the conclusions of a research paper should be based on the research contained therein, and not merely be the personal opinions of the authors about the general topic matter. Yet often peer-reviewed public health research articles feature the unsupported personal opinions prominently in the conclusions. Conclusions that do not follow from the research also include suggestions that specific policies are warranted (unless the research includes policy analysis, which was not the case for any paper in our dataset), general pronouncements about policy action (declaring *something* should be done with no exploration of the basis for making normative declarations), declarations that a problem exists when the research is incapable of showing there is a problem (e.g., the research shows that something is occurring, but cannot address whether it is harmful), and, obviously, statements about the study results that are not actually supported by the data. More subtle failures include most calls for further research (which might indeed be useful, but these conclusions are seldom related to the research results or any analysis) and specific conclusions based on apparent lack of awareness that knowledge exists beyond the study (e.g., if results of the research are merely what experts already were confident was true, they cannot be claimed to be the basis for new advice [Phillips: http://www.ncbi.nlm.nih.gov/pubmed/11511601]). Despite the papers in our dataset being rife with all of these errors, not a single journal reviewer challenged them. Sophisticated readers typically skip the "implications" passages of research reports because they are generally

worthless and misleading, but naive readers often read nothing else, so it is critical for journal review to correct such glaring errors.

Similarly, the reader should not be asked to accept summary assertions about results *in lieu* of reported results. The interview-based research papers in our dataset, as well as some of the others, were characterized by this problem. The reviewers failed to ask for some assurance that subjects' discourse was not summarized or selected to serve the authors' goals rather that reflect the results of the research, including in cases where it was evident this was true. Allowing such inadequate reporting of results and method may reflect a failure to recognize that this is not 1985, where places to put published words are at a premium. Researchers can easily publish, somewhere, entire survey instruments (necessary for adequate reporting of methods) and entire interview transcripts and other datasets (so that interested readers can see the actual study results rather than just the authors' summaries); for BMCPH and many other journals, they can do so in the form of supplementary files attached to the paper.

Other common problems included introductions and discussions that are made up of inappropriate random points, often polemical and almost completely useless, about the general topic area that had nothing to do with the research at hand. The functional equivalent would be if the present paper included a discussion about the value of THR.

Though we excused ourselves from the role of copyediting in our reviews, our reviews identified numerous specific points that could have been corrected by the authors with trivial effort and without changing any substance (without even acknowledging methodological flaws or altering personally-favored conclusions). For example, Hughes failed to report the calendar time of data collection in the abstract, an obvious and easily fixed flaw, but the journal reviewers failed to notice and no correction was made. Almost no such problems were identified by the journal reviewers. Since these are relatively minor concerns, we leave it to interested readers to see examples in the appendices.

We observed a strong association between the quality of the paper, as judged by our reviews, and the quality of the journal reviews, driven mainly by the worst papers getting the worst reviews. This was basically inevitable; we could analyze only papers that were ultimately published in the journal, and a bad paper is much less likely to be published if it receives high-quality reviews. A non-inevitable trend that is reassuring about our approach is that the worse the paper, based on our individual or collective general impression of it, the greater overlap of our important comments. That is, the vast majority of serious flaws, as opposed to minor suggestions or highly technical flaws that could slip past any single serious reader, were noted by two or all three of us.

A further educational observation is that for the higher-quality papers our comments had relatively little overlap, and this was also true for the comparatively minor points we noted for the papers with fundamental flaws. This illustrates that a multitude of readers is generally needed to identify the many potential incremental improvements in what is already a good paper. Even for points that most expert readers would immediately agree are valid upon being prompted, most will be overlooked by any a single reader working in isolation.

We are aware that the content of two of the papers had been subject to extensive review by numerous experts before being submitted to a journal. Burstyn circulated a working paper for months, collecting numerous expert comments, before finalizing the version that was published. Shepperd et al. had been developing their protocol for years before publishing it, working with a large team within their corporation, and bringing in outside experts for consultation (including CVP), though the manuscript itself might not have received such scrutiny. By contrast, we are not aware of any of the other papers or their core content being circulated for expert review before publication. Presumably not coincidentally, the only two papers that any of our reviews judged to be free of fundamental flaws were Burstyn and Shepperd.

We begin our analysis with those papers, and continue with articles grouped by study type. We believe that the Burstyn paper represents what the average reader assumes about the journal review process: A paper that has already been honed through scientific discussion is improved incrementally thanks to suggestions by reviewers. We further believe that the average reader assumes that other papers that include reparable serious flaws are substantially revised thanks to the review process, and those that are irreparable are never published in a journal. The vast majority of the papers in the dataset shows just how wrong these perceptions are.

Burstyn

Burstyn analyzed all available data about the chemistry of e-cigarette aerosol and reported that the quantity of exposure to levels are well below those estimated to be harmful. In consideration of our collective positive bias toward this paper, our reviewers (BLC and CB) were instructed to search especially hard to find something to express concern about; they identified only points of clarification and suggestions for extending the research beyond its existing scope.

The journal reviewers (PH and KF) also identified only points of clarification and suggestions for extensions. (Note that to minimize distraction, we designate the journal reviewers with their initials; their full names appear with their reviews in the appendices.)

Reassurance that none of these reviewers overlooked glaring flaws of the sort that appeared in most of the other papers in our dataset can be found in the history of the paper: It has been accessed 120,000 times at the BMCPH website and is often referred to as the single most

important paper about e-cigarettes, and we are aware of no published commentary offering substantive criticisms that were not identified by the reviewers in our dataset (except for one by the author -- see below). A prominent attack on the paper by an activist researcher, who found the results inconvenient and thus attempted to challenge them [http://www.tobacco.ucsf.edu/new-e-cig-risk-assessment-uses-wrong-standard], failed to identify any points of substance other than the observation that the author could have provided the context of other exposure limit standards rather than just than the one he used [http://antithrlies.com/2013/08/11/glantz-vs-burstyn-hardly-a-fair-fight/], a point also identified by our reviewers. (In general, when a paper is important enough that some people who work in the field are ideologically or financially motivated to dispute it, it offers us a useful epistemic tool: If they fail to identify flaws despite their best efforts, we have strong reassurance that the analysis -- at least the part that is reported in the paper and thus can be assessed -- is solid.)

The Burstyn review process illustrates a common popular misconception about peer review: The author accepted about half of the suggestions of the journal reviewers but indicated his belief that the other suggestions would make the paper worse rather than better, and did not make the changes. The suggestion about comparing the results to other exposure limits was made multiple times when the paper was circulated for comments and the author rejected it (arguing that because observed exposure levels were *so* far below any limit or any other accepted limit, it would add nothing). This general pattern is common for suggestions about details, whether offered by a designated reviewer for a journal or an interested reader of a working paper: the author often judges his original version to be better than the reader's suggestion. The journal implicitly accepted the author's reasons for refusing the suggestions and published the article which (in our obviously biased opinion) was reasonable in this case where the paper was solid and the author was expert. However, authors sometimes blindly accept journal reviewer suggestions that make the paper worse and journals often also let authors get away with brushing off serious bright-line flaws that reviewers identify; both of these phenomena were observed in our dataset. It might come as a shock for those who idolize journal publication that the content of a paper in a journal ultimately is determined by the professional opinions, skills, and honesty of the authors, as well as the limits thereof, and this not magically changed by the peer-review process.

As an additional lesson from this case study, it turns out that an astute reader identified a clear error in units in one Burstyn result that is now noted in a correction by the author [http://www.biomedcentral.com/1471-2458/14/18/comments] (this information did not exist at the time of the reviews for this project). The error did not change any general conclusion of the study and was part of a heuristic calculation rather than a real quantification, but it is still a bright-line error, and the author did change the intensity of one of his qualitative conclusions as a result. No one among those collaborating on the present analysis, the journal reviewers, or a

hundred thousand previous readers (including those actively attempting to denigrate the research) noticed the error. The post-publication review process finally caught and corrected the error, but this goes to show just how robust a process is needed to ensure accuracy of the scientific record (to say nothing of honesty and willingness to volunteer corrections on the part of authors), and that journal review is not remotely sufficient for that. The implications of this and a similar story, which have been recounted in more detail by IB and CVP [http://antithrlies.com/2015/06/11/post-publication-peer-review-correction-to-burstyn-2014-and-related-matters/], are arguably as important as any other observation in the present analysis for showing how misplaced the blind faith in the journal review process is.

Research protocols

Shepperd

Shepperd et al. provided a research protocol for a trial of reduced-toxicant cigarettes that will collect biomarkers of biological effect data from smokers who undergo a forced switch to the alternative product.

One of our reviews (BLC) expressed serious concerns about this paper; the others noted some of the same points of concern but were more positive overall. The main concern was that the justification for this human subjects experiment was inadequate, with the authors failing to present what reductions in effects they expected, failing to argue that these would be relevant to health outcomes, and largely basing their justification for the research on a single published opinion piece. In addition, we noted that the statistical analysis plans were largely unreported and somewhat dubious, and there was no justification for the claims of adequate statistical power. We expressed further concern that many other details were left out and that a few of the stated aspects of the protocol seemed destined to fail. BLC noted that the great demands imposed on the study subjects, along with the undisclosed level of compensation offered them, may create both threats to the study's internal validity and ethics problems, as well as questions about the representativeness of people willing to participate. One of us (CVP) was relatively unconcerned with most of these problems, being of the opinion that if all the data from such research is gathered and reported, then many of the details can be dealt with later; but his review noted that the paper could have been improved in several substantial ways, and did not dispute the validity of the greater concerns expressed by the others.

The single journal reviewer (YL) identified numerous details that the authors could have reported better, which were corrected over the course of several iterations of revision and re-review. YL's was one of very few reviews in our dataset that demonstrated deep technical knowledge about the study methods. However, he did not address questions of statistics or justification, or suggest there was anything that needed to be corrected or rethought.

This case study provides a useful lesson on how the journal review process is inherently inadequate, even when both the submission and the reviewers are generally high quality: While any competent reviewer ought to recognize bright-line serious flaws (as evidenced by the concordance of our comments on almost all other papers), credible reviewers may disagree about what constitutes a flaw in other circumstances. Our reviews noted concerns in areas that YL did not address at all. If one of us had been added as a second reviewer of this paper, the results would have depended substantially on which it was: CVP would have called for improving the stated justification for the work and challenged a few of the methods, but would have shared YL's lack of concern about the statistical methods, while IB would have called for a thorough revision of the statistical methods reporting and would have said less about the justification. BLC would have noted all of these and demanded far greater effort to justify the study and address its limitations. Whatever one's opinion about which of these views is valid, it should be clear that even if journal reviewers provide expert, careful, fully-thought-out reviews, important aspects of a final published article depended on which reviewers happened to be selected. This is even more true when selected reviewers might be inexpert and sloppy, but this case illustrates there is nothing magically objective about the journal review process even at its best.

Moreover, any second reviewer would have clearly been inadequate. All papers contain hundreds of details, and in this case there are far more than usual because of the myriad study methods. The four reviewers identified an almost non-overlapping list of needed clarifications and concerns at the detail level. A similar pattern was observed in the points of clarification about Burstyn. When readers are suggesting detailed improvements, anyone who has circulated a paper for comments can attest that new useful comments can still be found in the fifth, tenth, or twentieth review. But the journal review process seldom involves more than two or three readers, so even when the paper is good and the reviewers are skilled enough to offer any useful input, they cannot catch everything.

The reader can see a further illustration of this in Table 2. Recall that we included less-than-fundamental problems in the table only if we all agreed they were valid. Then note how many such issues were identified by one of us but not the others. Considering this as a sample-resample process, it suggests that each of us identified well under half the points that we all agreed were valid upon when one of us pointed them out. The fact that particular improvements appear in a final peer-reviewed publication, while other potential improvements do not, is an accident of fate resulting from the choice of reviewers.

Manzoli

Manzoli et al. presented a research protocol for a cohort study of e-cigarette use, focusing on behavior, but with some attempt to measure health outcomes. Its incompleteness and bright-line problems provide a striking contrast with Shepperd.

Most notably, all our reviews pointed out that paper failed to explain or even identify, let alone justify, the methods of the research, even though that was the entire purpose of the paper. Despite the inadequate reporting, we were able to identify numerous concerns about the research plan, including that the recruitment plan seemed to create serious selection bias, and finding flaws in the proposed statistical methods. In particular, the authors planned to assess health outcomes, but their methods seemed completely inappropriate (especially due to confounding by recent smoking). Even putting aside that general concern, the proposed study was grossly underpowered to measure health effects and there were specific errors in the plan. The paper also failed to report adequate background, and the discussion was pointless. It may be that by the time of the journal review, some of the flaws could only have been acknowledged but not corrected (e.g., the subject recruitment may have already been underway when the paper was reviewed – a problem in itself, though not relevant to the present analysis), but the failure to adequately explain methods could have been corrected in the manuscript and some flaws in the research plan could have been corrected.

The single journal reviewer, EF, identified only one minor point among the many failures to adequately report the methods. He made a suggestion about biomarker methods for detecting product use, as did one of our reviews, though he appears to have gotten the details wrong. His other comments were one trivial clarification and a call for adding a bit more irrelevant information to the discussion material. The net effect of the journal review process was perhaps slightly on the positive side, but completely failed to address the serious fundamental flaws in the paper, which remained in the article version.

Systematic behavior studies

Polosa

Polosa et al. conducted an intervention study of 50 Italian smokers, giving them open system e-cigarettes and educating them on their use. They observed that many quit or reduced their smoking. The authors characterized the subjects as not interested in quitting smoking and the intervention as not encouraging them to quit.

Our reviews included numerous observations of the study being described inadequately and seemingly inaccurately: The details of the recruitment are missing, leaving the reader uncertain about the target population and possible selection bias and casting doubt on the claim subjects were not interested in quitting. The details of the intervention were missing, creating serious doubt about the claim that subjects were in no way encouraged to switch to e-cigarettes. Our reviews also noted ambiguous and confusing reporting of the results, inappropriate reporting of statistics, dubious statistical methods, dubious interpretation of laboratory results, and use of undefined and badly chosen jargon. We noted that much of the discussion was based on unstated

or unexplored premises, and political bias was apparent in the results reporting. We also noted that the authors overstated the advantages of study design details and did not explore how this intervention related to real-world experience. In particular, the result is a huge outlier compared to real-world observations (particularly if we assume the subjects did not really want to quit smoking and the intervention did not encourage it), but the authors failed to acknowledge this or offer a convincing explanation. We also noted that many of the behavioral and all of the policy conclusions did not follow from the research.

One of the journal reviewers (JLH) offered no substantive comments. The other (HM) recommended different choices of what results to highlight and how to cut the data differently. In particular, this included favoring the biomarker measures over self-reported smoking cessation and offered some methodological concerns about their measurement, though did not note the concern about use of the wrong cut-point between positive and negative results that one of us did. He also questioned some of the use of terminology, though did not note the larger problems with it that we identified. Neither reviewer offered any comments suggesting they were bothered by the bold conclusions. The authors incorporated the recommended small clarifications, but rejected the more substantive comments and were allowed to do so.

Overall, our assessment was that this was interesting and informative research, but the data analysis and report, while not terrible, suffered from numerous ambiguities, errors, and overstatements that could easily have been corrected by a robust review process, as might be achieved by circulating a working paper. These flaws were left unchanged by the journal review process. This case study is a particularly good example of how the journal review process is no substitute for the more robust peer-review process of authors collecting the feedback of many readers with relevant expertise before finalizing a paper.

Joffer

The authors describe a longitudinal cohort study conducted in Sweden which observed 12- and 13-year-olds and followed up at ages of 17-18 with the stated goal of identifying predictors of smoking. Subjects were surveyed on tobacco use and a variety of intra- and interpersonal, behavioral, and familial factors as they may be related to future uptake of cigarette smoking. The authors concluded that low self-esteem, less negative attitudes about smoking, and ever use of snus were predictive of eventual cigarette smoking and recommended policy interventions. (This study did not explicitly address THR, but qualified for the dataset because it focused on the use of a THR product in the context of smoking choices.)

Our reviews identified numerous serious concerns about both the research and its presentation. The methods reporting was insufficient, but we still could identify serious concerns about the validity of the survey and the coding of the data. The authors consistently confused predictors,

which they claimed to be studying, with causes; most notably they suggested that snus use causes smoking, which is clearly not supported by their results and is contrary to existing knowledge. The results were inadequately and improperly reported in many ways. Even setting all that aside, the main conclusions did not follow from the results and the seemingly most interesting implications of the actual results were barely mentioned. We noted that most of the content of the introduction and discussion was useless.

One of the journal reviewers (SH) noted that the method for measuring the outcome of interest was inadequate, as did our reviews. The authors refused to make the change on the basis that a better measure would give them less statistical power; amazingly, the editors allowed them to get away with this. Neither journal reviewer identified any of the other serious problems in the survey, nor any of the problems with the data analysis or presentation, nor that the conclusions did not follow from the results. SH made a suggestion about the statistical methods (which the authors accepted) that arguably made them worse; in any case, it did not address the fundamental problems.

Some of the serious problems we identified could have been corrected with a rewrite, while others would require reanalyzing the data properly; without better reporting of the methods, it cannot be determined whether the irremediable problems with the survey are full-on fatal flaws that would constrain what could be analyzed. None of these problems were addressed by the journal review process. The journal reviewers identified various minor corrections and had the authors eliminate a lot of extraneous discussion, so the net effects of the journal review process were slightly positive. But the outcome was just a cosmetic improvement that left serious core flaws unchanged.

Hughes

The authors describe the results of a survey, given to 14-17 year olds in North West England, assessing the respondents' smoking and drinking behaviors and "access" to e-cigarettes. The authors claim that a substantial number of teenagers, including those that do not smoke, are accessing e-cigarettes. The authors further conclude there is an "urgent need" for controls on the promotion and sale of e-cigarettes to children.

Our reviews identified a common fundamental problem in survey research about e-cigarettes: that the key question on which the entire analysis was based was a single survey question that asked whether someone *ever tried* an e-cigarette, but the authors misinterpreted that as a measure of *using* e-cigarettes. Even worse in this case, the question was an either-or question (an affirmative answer could mean they had purchased an e-cigarette but never tried one) which is poor survey methodology. In addition, we noted that the survey sampling method was seriously problematic, making it effectively a convenience sample. The authors state policy conclusions

that in no way relate to the research. Their research merely showed the associations of behaviors we would expect absent any causation among them, and thus could not be the basis for concluding that any of them cause any other. Moreover, they do not even establish that causation would represent a problem. Yet their conclusions are all about fixing supposed problems. Moreover, to the extent that their results tend to support any causal conclusion, it is that that e-cigarettes are being used by teenagers for THR, flatly contrary to the authors' alarmist spin. The authors offered interpretations of their results that were purely speculative and often apparently wrong. In addition, we noted serious problems with the statistical methods and reporting, inadequate reporting of the research methods, and background unrelated to the study.

One journal reviewer (ZBT) said the authors did a "great job" with the research, apparently oblivious to the fundamental problems with the survey, let alone the authors' interpretations. He noted a possible selection bias, apparently not recognizing that the convenience-sample nature of the survey renders this moot. In addition, he recommended adding another speculative conclusion that is unrelated to the research. Both reviewers noted minor labeling problems in the reporting of the statistics, but failed to notice the serious errors in the reporting and statistical methods. The other reviewer (GK) asked for one clarification about what a particular survey question said, but failed to generalize that to the need to report verbatim all the relevant survey questions (especially in light of the fact that the main question of interest was so garbled).

This paper could have been turned into a legitimate, albeit relatively uninteresting, contribution to the literature by paring it down to a simple reporting of the survey results (correcting the statistical errors), with an explicit presentation of the severe problems with the survey, and no claims about policy implications. But the journal review process merely suggested a handful of cosmetic improvements and ignored serious problems. Notably, both reviewers identified small points that were aspects of broad fundamental problems, but saw only a few trees while missing the forest. The net result of the journal review process, apart from cosmetic fixes, was neutral at best.

Popova

The authors conducted a lab study of responses of nonsmokers to fictional warning labels for smoke-free tobacco products and wrote about that briefly within what was basically an extended political commentary advocating aggressive and misleading warning labels that was in no way supported by their research. As previously noted, our reviews of this (by CVP and CB) have been published as comments at the journal (this content appears in the appendix). These comments were based on the final version of the paper which had all the same fatal flaws as the original submission plus one major problem that did not exist in the submission because one journal reviewer asked for removal of a critical reported result and the authors complied.

Our reviews noted that the study was incapable of producing any result other than the obvious fact that when people are given scary or reassuring information about something, their degree of concern increases or decreases (the study design could only measure such qualitative change, not quantify it). We also noted serious concerns about human subjects ethics violations by the authors. The paper itself was almost entirely a political commentary rather than a report and analysis of the study. The conclusions were unrelated to any result from the study, but instead were based entirely on an unstated and indefensible premise that risk communication should always make people think there is more risk. We noted that the conclusions were not merely unsupported by the research but flatly contrary to the implications of the results: the results clearly showed that subjects overestimated the relevant risks, but the authors tried to hide this important result and called for more aggressive warnings. In addition, we noted that the methods were inadequately described and had problematic details (even beyond being inherently useless and unethical), and the introduction did not relate to the research.

The journal reviewers identified none of these obvious problems. One review (IA) was content-free, but that made it the better of the two. The other (SS) provided no comments that addressed the flaws in the paper, and actively praised the conclusions and introduction that were unrelated to the research. Worse, she advised the authors to suppress their results that most clearly demonstrated that subjects overestimated risks at baseline (and more so after seeing the misleading "warnings"), and the authors did so and the editors let them.

There is an argument that field research, once done, should be published in some form, if appropriately epistemically modest and not bundled with unsupported conclusions; there is no point in discarding even tiny bits of knowledge. But Popova appears to be one of the rare exceptions, given a study design that could not provide any useful information and the fruit-of-the-poison-tree ethical issues. But if the results were allowed to be published, the journal review process could have at least improved the paper in many ways, notably by not allowing the authors to report conclusions that were flatly contradicted by their study results. Instead, the journal review process actually made a horrible paper substantially worse.

Interview-based research studies

Edvardsson

The authors claim to be reporting theme-based results from a focus group study of 27 Swedish adolescents (apparently mostly of majority age, though the authors misrepresent this). But the paper reads like a series of temperance pamphlets with motivational quotes, rather than a scientific report.

Our reviews included some disagreement about the fundamentals. We agreed that methods reporting (which was inappropriately scattered throughout the paper) omitted key information,

but we differed in our ultimate assessment. BLC, who has done more using such qualitative research methods, found them generally adequate for this type of research but CVP and IB felt the methods reporting was almost completely uninformative. CVP characterized the results reporting as more like feature-story journalism, with the authors telling a story of their choosing and embellishing it with quotations, rather than scientific reporting of study results. CVP and IB said that what results were reported seemed like nonsense. Were it not for our requirement to review the details despite fundamental flaws, CVP and IB would have insisted on a major revision before trying to analyze the paper. BLC was somewhat less critical, but still noted considerable conceptual problems with the presentation and interpretation of the data, which demanded immense trust on the part of the reader. (Note: For those who may not know, this comparison to feature journalism reflects a perennial debate about the scientific legitimacy of this "qualitative research" style, wherein the reader is forced to accept the authors' summaries and assertions to an even greater extent than with other research. For an illustration, imagine the present paper, but without Table 2, the Appendices, or the systematic description of the content of the journal reviews.) We agreed that the introduction failed to address key concepts and background, while containing pointless and incorrect claims, and that the authors seemed unaware of basic facts about snus. We agreed that the authors' categorization and their key concepts were based on vague hand-waving rather than clear definitions.

This appears to be the only paper in our collection that went through a full-on revise-and-resubmit process with two rounds of reviews. The journal reviewers were generally positive about the first version. In the first round, one reviewer (BC) offered observations about the methods and results reporting that collectively could be interpreted as agreeing with us that there was a broad and profound failure to report the methods. However, his tone suggested that these were minor problems and he did not indicate what corrections were needed. TP generally praised the study, expressing concern only about the sample that was recruited and about the conclusions about male-female comparisons given that few women and girls were studied, though he then said to leave it in anyway. NW explicitly said that the methods and results were reported adequately, and mostly called for the addition of more vague framework theories. None of the reviewers noted any of the factual errors (see the discussion of choices of reviewers, below).

The revised version was tidied up a bit, but the substantive content was unchanged. The factual errors and failure to report proper background were not fixed, as would be expected given that the journal reviewers did not note the problems. But the authors also ignored most of the valid criticisms from the journal reviewers. The two reviewers who provided second reviews (BC and NW) seemed unbothered by that and focused entirely on copy editing. The final published version was trivially different from the original submission despite the additional round of reviews. Someone reading the original submission and then the final article would have a

difficult time identifying what non-cosmetic changes were made. While our views of the initial version ranged from likening it to the Sokal hoax to it merely suggesting substantial improvements in the reporting, we agreed that it required major changes, but the journal review process accomplished little more than style and copy editing.

Atkinson

The authors conducted face-to-face interviews of 36 "disadvantaged" smokers and former smokers in the UK to explore the possibility of encouraging the temporary substitution of NRT when in the home, to reduce children's exposure to smoke. The authors found almost no interest in the possibility. Their interpretations of the results attribute this to subjects' perceptions that NRT is for smoking cessation and a lack of understanding about the value of temporary abstinence. Nevertheless, the authors conclude that using NRT for this purpose is still a promising approach.

Our reviews included numerous observations that the authors' were mostly reporting their own premises and personal views of how people should act, rather than listening to the subjects and reporting the resulting data. We noted that the main conclusions did not follow from the results but simply reflected the authors' preconceived assumptions. In particular, the study results suggested that this strategy offered little promise. The authors did not explore the reasons for this, but rather asserted the opposite conclusion. The reported conclusions that did follow from the results were uninteresting, but we noted several interesting implications of the reported interview data that the authors do not even mention, let alone explore. This is presumably because these results did not conform to the authors' premises. The obvious biases colored the results reporting and dominated the discussion, reaching the point of being overtly patronizing toward the subjects. We also noted that the reporting of the results was mostly in the form of interpretation, rather than presenting the data; while not the near-caricature of qualitative research reporting of Edvardsson, it still forced readers and reviewers to simply accept the feature-story-style assertions of the authors and assume the illustrative quotations were representative of a broader pattern. In addition, we noted that the methods were inadequately explained, the introduction was missing important background, the study did not seem grounded in existing knowledge or theory, and various specific claims in the results were not supported.

The two journal reviewers identified none of the glaring problems. ARH mentioned some generally related studies that were missing from the background. AP's review had no substantive content.

We observed that this research had potential to offer some useful information, and a robust review process could have brought that out. This would have required eliminating the conclusions that were unrelated to, and even flatly contradicted by, the study results and

replacing them with better reporting of the data and exploration of what it showed. Instead, the journal review did nothing to improve the paper and the inadequate reporting, destroying all potential value of the effort, which creates ethical concerns about the imposition on the human subjects.

Silla

The authors conducted brief telephone interviews of 15 university students, divided into 3 groups by smoking and NRT use histories, with the goal of investigating the potential for use of NRT for tobacco harm reduction. They identified 24 themes, focusing on 3 in their discussion, and concluded that past NRT use increases "engagement" and that participants had significant misperceptions.

Our reviews noted the limitations of the small sample, unusual population, biases among the study subpopulations, and the brief interviews (30 minutes to explore 24 themes). The authors were not properly epistemically modest about their results given these severe limitations. The methods were inadequately reported, though probably better than average for studies of this type. We observed that the emphasized conclusions did not follow from the study results, and sometimes flatly contradicted them, while apparent important implications of the results were glossed over or ignored completely, including those that relate to the supposed aim of the study. Several of the conclusions were based entirely on the premises of the authors – all unexplored and some clearly false – rather than anything in the data, and the tone became rather derogatory toward the subjects. We observed that the authors seemed to be hobbled by their premises and misconceptions: they misrepresented the effectiveness of NRT, seem to not understand that NRT marketing contributes to misperceptions about those products, often confused THR and abstinence, and ignored the benefits of smoking and nicotine in their analysis (a crucial point). As a result, they failed to ask subjects what attributes of NRT might make THR more appealing. The reporting of quantitative results was completely inappropriate for the small convenience sample, and some of the quantitative claims were false.

One of the journal reviewers (GG) offered no comments and stated that the submitted manuscript should be published unchanged. The other review (AD) was the most useful journal review in our dataset, in terms of trying to correct some of the core flaws in a submission. AD noted that there were "fundamental flaws" in the analysis and identified some, but far from all, of the flaws we did as well as a few valuable details that we did not. She also suggesting adding some irrelevant points. AD noted the odd study population; the authors added a throw-away acknowledgment but did not make their conclusions any more modest. AD expressed doubt about the subgroups of five reaching thematic saturation, as did our reviews; the authors merely added an unsupported assertion that no new themes emerged in the last interviews.

In spite of AD's review being the best of the journal reviews that we analyzed, it was still relatively brief and not adequately clear with some of the criticisms, resulting in the authors not making the important changes. For example, she made the general statement (in agreement with our reviews), "Several places statements [sic] in results are either conclusions or go well beyond the data presented and the goals of the study", and offered a single example. The authors responded only by removing that one example, with no attempt to correct the general problem and many other examples that we identified. There is no record of the editor offering AD the opportunity to note that the single change did not respond to her general point. Similarly, AD (in agreement with our reviews) stated, "Most importantly the authors stray in their results/discussion/conclusions from their focus on their stated purpose." The author protested that there were no specific instructions contained in the observation, and so ignored it. AD noted that a few selected quotations were represented as if they were universal, that the methods were inadequately reported, and that many conclusions did not follow from the results. The authors' responses to these were trivial changes that did not address the problems. Additionally, despite identifying some of the fundamental flaws, AD overlooked several points that our reviews agreed were glaring errors.

Because of AD's contribution and some potential positive value of the research project, this paper appeared to be the best candidate in our dataset for the journal review process to fix a seriously flawed paper. But instead it served only to further illustrate that the process fails not just because of poor reviews, but because of the process itself. Journal review only works if an assigned reviewer is capable of noticing the serious flaws in a paper, at least a substantial portion of them (a unique occurrence in our dataset), *and* is given a chance to make sure that these flaws are corrected. The mere act of allowing a reviewer to note the flaws but then publishing the paper with the flaws still in place renders the process worthless even if a skilled and dedicated reviewer is recruited. This observation (in contrast to the observation that reviews tend to not be high quality) does not necessarily generalize across journals or editors – another might never allow comments like AD's to be effectively ignored. However this does demonstrate that some peer-reviewed public health journal articles are published in spite of containing fundamental flaws that a reviewer identified.

Literature review articles

Zyoud

Zyoud et al. conducted a simplistic exercise in which they searched a single database of articles for papers relating to e-cigarettes and reported some counts (e.g., of how often someone was an author of such a paper). They claims various grandiose study aims.

Our assessment was that this was an utterly pointless exercise that could not possibly address the study aims and, indeed, had no apparent value whatsoever. We further noted it used a flawed

search strategy, provided none of the useful analysis that should have been included, and stated bold conclusions that did not at all follow from the research methods. In addition, we noted that it implicitly perpetuated the myth that most of the useful research on e-cigarettes can be found in journal articles, overstated what it found by conflating commentary with research, and recited incorrect claims about the subject matter that were unrelated to the research.

Two of the three journal reviews (MRG, PC) were null-content, simply recommending it be published. The other reviewer (KF) identified only one of the several failures to analyze the content usefully that we identified, and two specifics among the many errors in the tangential commentary. The authors responded to the former with a trivial change that did not actually address even that specific point, and the editor and reviewers allowed that to stand. KF also asked that the time covered by the review be updated, a point addressed below.

The only apparent way that the journal review process could have salvaged this paper would have been to instruct the authors to scrap it and do a useful analytic literature review instead. Failing that, the entire content could have been condensed to a short note with one paragraph of methodology and two tables; that would have been basically useless, but at least it would have been accurate and harmless. Instead, the review process did not even cause the authors to eliminate the inaccurate commentary or completely unsupported conclusions.

Luo

The authors conducted a sort-of-systematic review of YouTube videos that portrayed or mention e-cigarettes, and coded their content. However, this seemed to just be an excuse to write a political commentary; the authors offered no suggestion of what scientific value their research might have.

Our reviews noted that the conclusions did not follow from the research and many of the authors' premises were false. The study methods were inadequately described, but there was enough information to know that they were fatally flawed. Our reviews noted that the coding seemed arbitrary and purely political and thus the results were meaningless. In addition, the introduction and discussion were pointless and contained much that was wrong. The entire analysis was based on false premises about e-cigarettes and unexamined (and almost certainly false) premises about the behavior and beliefs of YouTube viewers. On top of that, one of us noted that the authors did not even seem to understand how YouTube searches work, and another noted that different search strategies yielded radically different results. (These latter observations were was based on no extant expertise, but on the simple expedient of opening YouTube and running a few searches, an action that could be performed as part of even a cursory review.)

Of the two journal reviews, one was vacuous (LM). The other reviewer (AA) noted two of the relatively minor issues with the methodology (which were among the many noted in our reviews), but the authors refused to make any changes in response. AA tried to tinker with the introduction and hinted that the assessments in the discussion were invalid, but the authors made no substantive changes. AA suggested that the authors add even more irrelevant material about anti-tobacco politics; this change was made, while his slightly useful comments were ignored.

To salvage this worthless research and the negative-value paper, the journal review process would have had to instruct the authors to identify some possible justification for the research, learn something about what they were studying, redo the research completely, eliminate the factual errors, and report research results rather than political polemic. Instead, the review process actually managed to make a paper that already had substantial negative value even worse.

Choice of reviewers

A systematic analysis of the identity and skills of reviewers was not part of our research. However, it was impossible to not be struck by the observation that, with only a few exceptions (the Shepperd reviewer, the Edvardsson reviewers, one Silla reviewer), it appears that all of the reviewers were chosen based on having written in the general topic area, not because of expertise on the particular type of research being reported. This is typical in public health publishing, and is often not too problematic due to substantial overlap between subject matter and research approach.

But not always. Most notably, all three reviewers of Zyoud were clinical researchers with no apparent expertise in bibliographic analysis, though there is no reason why a reviewer of that paper would need to know anything about e-cigarettes or health science at all. Similarly, the reviewers of Luo had no apparent expertise in media analysis. The reviewer of Manzoli had no apparent background in social science methods. These papers were seriously flawed, but the reviewers expressed little concern. Research using other than standard public health science methods (i.e., anything other than epidemiologic field studies, toxicology, or particular aspects of socio-psychology) needs reviewers with specific expertise in the methodology. The pattern of selecting reviewers familiar with the subject area but not the research methods is particularly ironic given what the authors chose to try to explain. In the papers reviewed, as is typical, many of the authors attempted to explain details of the subject matter that were irrelevant to the research, such as explaining how an e-cigarette works. However, attempts to explain research approaches that might not be familiar to many readers are missing (in our data, absent from all but Burstyn). Such explanations are necessarily limited – it is not possible to explain survey methods or what a gas chromatograph does in a research study – but it does mean if none of the reviewers are expert in the methods themselves, then important errors might be overlooked.

On the other hand, reviewers with background only in the particular methodology alone are also not so useful. The Edvardsson reviewers apparently all came from the particular behavioral science tradition that employs the reporting style from that paper, which some of us find inadequate. They seemed to fill in the blanks for some of the unexplained methodology jargon, but the result of this is that they did not call for explanations of methods that are obscure for readers of a public health journal. They were oblivious to the subject-matter factual errors. Since the paper was being reviewed for public health sciences journal and makes conclusions about public health policy, a reviewer with such expertise was clearly needed.

It turns out that in this dataset, none of the fatal flaws and very few of the fundamental flaws identified by any of the reviewers required arcane knowledge or deep expertise. They should have been identified by anyone familiar with scientific reasoning, methods reporting, sample selection, and other basics. Even the more technical fundamental flaws we identified – e.g., the inappropriate trend analysis in Hughes – would at least be understood by a nonspecialist once pointed out.

Competing interests

BMC policy calls for authors and reviewers to disclose their competing interests. To BMC's great credit, this is not limited to the relatively uninformative reporting of who funded the research, but rather authors are explicitly instructed:

> Are there any non-financial competing interests (political, *personal*, religious, *ideological*, academic, *intellectual*, commercial or any other) to declare in relation to this manuscript? If so, please specify. (emphasis added; [http://www.biomedcentral.com/bmcpublichealth/authors/instructions/researcharticle])

One of us (CVP) was tasked with trying to identify undisclosed competing interests. He identified clear ideological/personal/intellectual competing interests among the authors of Popova, Hughes, Luo, Polosa, Edvardsson, and Atkinson (details in appendices, at the end of each CVP review) that were not disclosed by the authors. There were also a few cases of undisclosed financial conflict of interest.

BMCPH reviewers are instructed, "Whilst we do not expect reviewers to delve into authors' competing interests, if you are aware of any issues that you do not think have been adequately addressed, please inform the editorial office." [http://www.biomedcentral.com/bmcpublichealth/about/reviewers] We may safely surmise that in the several cases where one or more authors were well-known advocates in the area – either supporters of THR or opponents – expert reviewers would have been aware of this. In some

cases, this conflict of interest is patently evident from the content of the manuscripts themselves. Yet no journal reviewer addressed the point.

In addition, many of the reviewers themselves had clear financial and non-financial conflicts of interest, *none* of which were disclosed, though they have the same instructions to disclose as the authors. Both reviewers of Popova were employed by an organization that is ideologically committed to the conclusions in the paper, and one of them made clear her ideological biases in her comments. Two of the reviewers of Zyoud were, in effect, subjects of the study and stood to benefit professionally from its publication. Indeed, the primary comment from one of them was that the authors should expand the calendar time covered by the search; this resulting in including more of the reviewer's own articles in the data and thus substantially increased his own ranking in the results, something he presumably realized would occur. The failure to disclose obvious competing interests in that case study are particularly stark. Other examples are observed in the many reviews where reviewers were invested in the political conclusions of the authors and failed to note serious problems, particularly that the political conclusions in no way followed from the reported research.

In short, both authors and reviewers blatantly flouted the conflict of interest disclosure requirements. In many cases, no reference to the details of the journal's disclosure rules should have been necessary for them to figure out the need to report; the individuals would have no difficulty recognizing they had competing interests.

Politics and ignorance do not explain most of the problems

It is tempting to attribute many failures of journal peer review to pure conflict of interest: the reviewers and editors allowed bad papers to be published because they like the political conclusions they contained. But this is a facile explanation because papers can generally be improved without removing the political commentary or unsupported conclusions (though obviously they can be improved more by removing those).

Hughes et al. could have still attached their *non sequitur* policy conclusions to a paper that corrected the errors in statistical methods, and they could have at least reported the year the data was collected. Luo et al. could have attached their polemic to a valid review of popular videos rather than a flawed and useless one. The authors touting the potential of NRT could have fixed their methods and results reporting without removing the conclusions that did not follow from the research. (The exception is Popova where literally the only meaningful result of the research contradicted their political preferences; it is difficult to see how legitimate reporting of the study results was an option given the paper's political goals.) This pattern is easy to observe in the public health literature as a whole, not just our dataset. Even in the worst-case conflict-of-interest

scenario, where authors and those involved in the review process all wanted to publish political conclusions, it would still be possible to fix other flaws in the papers.

Similarly, the worst failures of journal peer review cannot be attributed to limited technical expertise on the part of the reviewers. Of course, those with naïve faith in the process probably believe that deep expertise is always present. But even a reviewer with rather limited skills ought to be able to recognize that a conclusion is completely unrelated to the research or that the methods section leaves him wondering what the study methods were. Little more skill is needed to recognize that important results are not reported or that the characterization of the results is not supported by the data. Those reviewing survey research should be able to recognize serious flaws in the survey design and all reviewers should have a basic knowledge of how to report statistics properly. We anticipated that during this research we would draw upon technical expertise we acquired over the course of our careers to point out subtle serious problems. There certainly are such problems in the public health literature, where a somewhat subtle study bias or modeling choice renders an analysis invalid. But the serious problems in this case series, as with much of published public health literature, turned out to be all at a level that we would expect to be evident to any competent second-year graduate student, and in most cases, to an attentive layperson.

Implications for faith in the journal peer-review process in public health

It is clear from this review that faith in the public health journal review process is misplaced. In our dataset we observed numerous serious bright-line flaws which should have been obvious to any skilled reader. None of them were corrected before the paper was published, including in the rare cases where a journal reviewer even identified them. Thus, except for cosmetic improvements, the reader would have lost nothing by just reading a working paper version of the manuscript. Indeed, in some cases the journal review process made the paper worse. The journal review process has inherent limitations that guarantee it falls short of the idolatry it receives (e.g., no matter how much skill and effort reviewers offer, they still only have access to the same paper that other readers do). But the reality is so much worse than the theoretical optimum.

Judging by length of the reviews received by the journal, most of the reviewers appear to have spent no more than a few minutes on their reviews beyond the time it took to read the paper. Indeed, many of the reviews do not even provide evidence that the reviewer read anything other than the abstract. Only one of the journal reviews was more detailed than the *least* detailed of our reviews of the paper (which were intended to be what we would have written in the role of journal reviewer). These observations alone point out the folly of the naïve beliefs about the review process: Presumably those with faith in the process assume that journal reviews are always as substantive as those we wrote, perhaps more so.

Many of the journal reviews offered suggestions for small improvements, of the sort that readers typically offer if a paper is circulated for comments, though many did not offer even that. In all cases, such suggestions clearly fell short of saturation, given the limited overlap of suggestions among all the reviews. This illustrates the obvious point that journal review, even if it were done right, is no substitute for collecting comments from many expert readers. Our experience conducting this study further illustrated the value of an interactive process, as is found in blogs or working paper comment systems: The individual reviews we wrote in isolation, as is typical for the journal review process, while far better than the journal reviews, were markedly inferior to the collective wisdom we could later assemble by triangulating our observations.

But good suggestions alone are not enough, whether made by journal reviewers or other readers. The value of the suggestions is mediated by the skills and honesty of the authors. We observed that for the seriously flawed papers in our dataset, bad suggestions by journal reviewers were about as likely to be accepted or rejected as good ones.

Post-publication review is potentially more effective at naming and shaming authors of bad papers, creating an incentive to avoid publishing material that is worthy of ridicule. Unfortunately, that "post-publication peer-review" process (a term which misleadingly implies that this is somehow an afterthought, when it is actually the bulk of the scientific process), a robust check on bad research in many fields, is almost absent in health sciences. The sheer volume of articles makes it difficult, if not impossible, for individual articles to receive additional scrutiny and discussion, forcing even experts to blindly trust most of what they read. The letter-to-the-editor process is wholly inadequate, restricting analyses to impossibly low word counts which are routinely rejected by the journals despite identifying serious or fatal flaws, and they are seldom even noticed if they are published. Health science journals make it very difficult to publish critical reanalyses or deconstructions, so the literature consists of a series of contradictory monologues with only just-so stories offered to explain the contradictions, if they are acknowledged at all.

Systematic reviews and meta-analyses are often conducted, but these rarely assess the quality of the research and tend to accept all reported results as legitimate. (This is apart from the methodology being inherently problematic in public health research.) These exercises can be done with more quality control and critical analysis [e.g., Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. Biostatistics. 2001 Dec;2(4):463-71. PubMed PMID: 12933636.EF], but in most current practice they just exacerbate the problem, codifying published results despite their contradictions and flaws and discouraging critical post-publication analysis. Stories like that of the Burstyn paper, where a reader pointed out an error to an author and the author endeavored to publicize it, are so rare in public health that it is difficult to find examples.

Moreover, communications of post-publication reviews are generally ignored in public health. A claim in a published peer-reviewed paper may be thoroughly rebutted in expert forums (e.g., blogs, conferences), but the rebuttal is likely to never appear in a health science journal except perhaps as a passing mention that will almost certainly be overlooked. This would not be a problem in fields where the consumers of the content are other experts who are familiar with the broader discussion, but it is a problem in health science where most readers are non-experts who are unable to discern the quality of an analysis and are unaware of the larger literature that exists outside of journal articles. Naïve non-experts (including clinicians and policy makers) are likely to never see the expert analysis, and blindly accept the conclusions stated in whatever journal article they happen to find. This amplifies the importance of the failures of the pre-publication review process. Worse, readers in public health are likely to cherry-pick or be directed to a journal article whose conclusions best align with their biases, and the most extreme conclusions are likely to come from flawed papers.

Of course, authors are just responding to the incentives created by the journal system. The lack of a post-publication review process, and particularly the lack of a market for critical reanalyses, leave little incentive for authors to worry about including blatant errors in their publications. The tangential wandering commentaries and irrelevant introduction are a result of journals allowing, or even encouraging, research reports to read like feature stories. There is no good outlet for reporting study results alone, without random bits of background or speculative interpretation. Of course, this does not excuse authors from publishing claims that are not supported by their data, but it does explain why they have the urge to write pointless, irrelevant, and speculative claims.

Some journal reviews are as thorough as the ones we created for this study (tautologically, since they were meant to be what we would write for a journal review), but many are not. Thus, all papers -- brilliant, good, bad, or terrible -- will almost certainly be published in a peer-reviewed journal if the authors want them to be. The fact that they are published in a peer-reviewed journal, therefore, tells us little about their quality. Authors of even patently flawed papers can be confident that a few submissions to different journals will eventually result in them drawing only low-quality reviews like most of those in our dataset, written by reviewers who either do not understand the flaws, do not spend enough time to find them, or intentionally ignore them because they like the conclusions. This includes cases of politically biased inaccurate analysis being attached to what could have been a somewhat informative study (e.g., Hughes, Joffer) and even cases where such political commentary is attached to a worthless study (e.g., Popova, Luo, Edvardsson). The stochastic process that often results in drawing favorably-biased or incompetent reviewers can be affected by the journal editor, of course, who can recognize reviews as inadequate and recruit new ones, or just recruit good reviewers in the first place. This means that what is observed at one journal does not necessarily apply to all journals, though not

too much should be made of this observation since it merely adds one more coin flip to the mix. The journal editor is just one more reviewer, and if he has the same biases or blind spots as the external reviewers, the same results will occur.

Moreover, it makes no difference whether *some* journals genuinely do a better job of gatekeeping: The public health journal review process, *taken as a whole*, performs no useful gatekeeping function. If authors find that entry into the peer-reviewed literature through a particular gate would require fixing the flaws in their paper that they do not want to fix, there are a hundred other gates, and it will not take long to find one that is unguarded. The majority of the papers in our dataset represent clear cases of a bad paper slipping through an unguarded gate. Publishing a paper in a public health journal is a game in which persistence, rather than quality, assures success.

The problem is further exacerbated by a "market for lemons" [http://qje.oxfordjournals.org/content/84/3/488.short] or "Gresham's Law" type problem for reviewers in this field: Writing a good review is not likely to accomplish anything, and so many would-be good reviewers refuse all invitations to write reviews that do not come from an editor who is a trusted colleague, providing some assurance he will take the review seriously and perhaps repay the favor. Many others refuse specific invitations when, upon previewing the material, judge it to be unsalvageable without a major rewrite. This further increases the proportion of reviews that are cursory or are just based on wanting to see the particular conclusions published. Those of us in public health who might be inclined to write a high-quality review for a journal, out of a genuine desire to improve the quality of the scientific record, have little incentive to bother. We all know from experience that the most likely reaction to a review that identifies major but correctable flaws is rejection of the paper, followed by it publication, almost unchanged, by another journal that solicited only low-quality reviews. This generally occurs even if the reviewer asks the editor to request a revision rather than reject. If the editor complies with that, the authors will often just withdraw the paper and seek an unlocked gate – after all, if authors have not circulated a paper for comments before submitting it, they are presumably just interested in getting it into a journal, not making it better. The next-most-likely alternative is what was experienced by reviewer AD with Silla: she identified fundamental flaws that could have been corrected, but they were not corrected and the paper was published largely unchanged. It is little wonder that she did not bother to go into more detail about her concerns.

Keep in mind that the failures we identified with the journal reviews of most of these papers are measured not against a gold standard, but against the problems identified in two or three reviews by individuals who undoubtedly missed some additional points of concern. We are confident that our specificity, while imperfect, is quite high, particularly at the level of the summary points we noted in Table 2 and the text. A few observations in our reviews were judged to be incorrect by

the other reviewers, but we believe it is extremely unlikely that other experts would dispute the general messages we emphasized in our broader analysis. But our sensitivity was undoubtedly well short of perfect, particularly at the detail level; we are unlikely to have overlooked a general failure to adequately report methods or that conclusions generally did not follow from the research, but the absence of a specific important methodological detail or *non sequitur* conclusory statement might have escaped us. Indeed, given how bad most of the papers were on these counts, forcing us to summarize our concerns, they inevitably did.

The sample from this study is not sufficient to suggest any quantitative generalizations. However, it is sufficient to demonstrate that many public health articles are published without any useful reviews and contain major flaws of various types that are obvious to many readers. Some journals in the field might have insisted on more complete or expert reviews before publishing, but BMCPH is a legitimate and respected journal, not some disreputable outlier – no one ever says "that paper is not credible because it was published in *BMC Public Health*." Articles on other public health topics might generate fewer reviews that are purely political; a large portion of publications on e-cigarettes, in particular, are thinly veiled political commentaries (something that one would not learn from reading Zyoud), possibly published because the reviewers and editors shared the author's politics. On the other hand, the reviews in our data also failed to note flaws that could be remedied without affecting the political conclusions. Reviews in other areas are perhaps less likely to be vacuous because the methods are more uniform and thus potential reviewers can at least identify glaring departures from standard practice, but since the standard practice is often quite flawed this helps little. Risk-factor epidemiology articles would usually be reviewed by someone familiar with doing risk-factor epidemiology, but this does nothing to stop the nearly ubiquitous problems of policy conclusions that do not follow from the research, unreported multiple hypothesis testing, naïve approaches to controlling for confounding, and so forth. Our experience suggests that our case series was perhaps worse than average, but not an outlier.

Conclusions

For the papers in our dataset, the journal review process contributed basically nothing to the scientific process. Papers with no value whatsoever were published. Studies with potential value were published with fundamental flaws that eliminated much of the potential value. In the very rare cases where journal reviewers identified one of the fundamental flaws, they were not corrected. Conclusions that did not follow from the studies were not removed. The only benefits from the journal publication process accrued to the authors, giving them a new line for their CVs and greater (and usually unwarranted) visibility and credibility. If the authors had merely published their submitted drafts as working papers, their scientific contribution -- or lack thereof -- would have been the same. Even knowing in advance the flaws in public health publishing, we were astonished by just how bad the observed results in our study were.

A paper published in a health science journal is what it is, but nothing more: It is the work of one or more authors. If they are good scientists, skilled in the methods, careful, and honest, then it is probably accurate and informative, but if they are not, it is quite likely not. It is up to the reader to assess the paper, or to decide to just trust the authors' skills and honesty. There is no alternative or shortcut. Trusting the journal review process is clearly a mistake. The journal review process ought to at least screen papers for being basically credible, but the results of our study show that even this is not the case.

In addition, all but the most naive readers know that articles are frequently retracted, sometimes quite spectacularly, for problems that are inexcusable but that no reviewer could be expected to ever detect, and the retractions are presumably just the tip of the iceberg. More important, a large portion of articles is simply found to have produced incorrect results even if they were an honest attempt at truth-seeking by skilled authors.

In public health, at least, the phrase "according to a peer-reviewed journal article" should be interpreted by the reader the same way they would interpret "according to a paper" or "according to a blog post", and should be given no greater deference: All those statements merely mean, "these particular authors are claiming this; if you trust their skills and honesty, then you should consider this informative."

Acknowledgments

The authors thank Clive Bates for his contribution to the research and suggestions on the manuscript, and Susan X Day for comments on the manuscript. CVP thanks Imperial Tobacco for support for this work in the form of an unrestricted (except for general subject matter) grant in support of research on peer review in public health; the funders were not aware of the specific project before a version of this paper was published. IB and BLC volunteered their time.

Competing interests

One or more of the authors are friends or colleagues with authors and reviewers of some of the papers, notably including Polosa, who is a member of the CASAA Board of Advisors. We did our best to not let this influence us in the spirit that we believe that the best compliment we can offer a friend or colleague is take their work seriously, and do them the favor of offering the best possible analysis -- they are, after all, our true peers. The authors are all positively disposed toward THR, and thus our reviews were probably more emphatic in their objections to anti-THR polemic, bias, and disinformation than the relatively rare pro-THR examples, but we did note those also. See also Acknowledgments.