28-Sep-2019
BMJ-2019-051528 entitled "Very rare pathogenic genetic variants detected by SNP-chips are usually false positives: implications for direct-to-consumer genetic testing"

Dear Dr. Wright,

Thank you for sending us your paper. We sent it for external peer review and discussed it at our manuscript committee meeting. We are interested in proceeding with it, and hope very much that you will be willing and able to revise your paper as explained below in the report from the manuscript meeting, so that we can make a final decision about publication.

Please remember that the author list and order were finalised upon initial submission, and reviewers and editors judged the paper in light of this information, particularly regarding any competing interests. If authors are later added to a paper this process is subverted. In that case, we reserve the right to rescind any previous decision or return the paper to the review process. Please also remember that we reserve the right to require formation of an authorship group when there are a large number of authors.

When you return your revised manuscript, please note that The BMJ requires an ORCID iD for corresponding authors of all research articles. If you do not have an ORCID iD, registration is free and takes a matter of seconds.

Thank you for entrusting us with your work.

Sincerely,

Elizabeth Loder, MD, MPH

*** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/bmj?URL_MASK=90d41123f25e4a579031c2a43c8e9d38

**Report from The BMJ’s manuscript committee meeting**

These comments are an attempt to summarise the discussions at the manuscript meeting. They are not an exact transcript.

Present: Elizabeth Loder (chair); Jon Deeks (statistician); Tiago Villanueva; Jose Merino; Wim Weber; Tim Feeney; David Ludwig; Helen Macdonald.

Decision: Request revisions following statistical report from Professor Deeks

Detailed comments from the meeting:

* We didn't think it was surprising that the lower the allele frequency the worse the SNPs perform, but thought it is useful to remind everyone about the clinical implications of this fact. One of our editors who works as a GP said "I am increasingly being asked about genetic testing and often struggle with what exactly tell patients."
* We think that many concepts could be better and more clearly explained. Please consider providing additional boxes with definitions and explanations of terms and perhaps even illustrations. This might lend itself to an infographic and perhaps a patient summary.

**Issues to include in boxes:**

**Definition:**
- SNP
- SNP-chip
- DNA sequencing
- NGS
- missingness rate
- Hardy Weinberg equilibrium also, explain why it is important
- SNV

**Illustrations (or further explanations)**
- How do SNP chips work and how they differ from NGS.
- Why are SNP chips good for common but not for uncommon variants

**Issues around generalizability of results from UKBB.**

* How often are commercially offered genetic tests based on SNP chips vs. other methods? Can physicians and patients easily identify which one is used in each particular offering?

* Figure 1 has results from reference 10, not from this analysis. This is confusing.

* We agree with one of the reviewers that the issue of Type II error should be addressed, which could be equally important from a public health perspective.

* The title seems more appropriate for an editorial and should not announce the findings. Please aim for a more neutral, descriptive title.

* Please try to do a better job at explaining the differences between SNP array (genotyping) and DNA sequencing. You have Fig. 1, but for the non-expert that is not very informative.

* Please see the separate statistical review by our head of statistics, Professor Jon Deeks. Please pay special attention to this in revising your paper. If his recommendations are at odds with any others from reviewers or editors, please default to following Professor Deeks's recommendations.

Please also revise your paper to respond to all of the comments by the reviewers. Their reports are available at the end of this letter, below. In your response please provide, point by point, your replies to the comments made by the reviewers and the editors, explaining how and where you have dealt with them in the paper.

**Comments from Reviewers**

**Reviewer: 1**

**Comments:**
The experiments reported in this paper appear to be well-executed and are extremely important: The likelihood of a true positive result using SNP-chips in direct-to-consumer genetic testing, "reduces substantially with decreasing allele frequency." It deserves publication with high priority. Unfortunately, the description of the experiments will leave many practicing physicians, whom I presume to be the primary target (Why else submit to BMJ?), befuddled. The fundamental problem is the authors’ failure to distinguish what SNPs tell us about disease risk from what gene sequencing tells us. This could be easily rectified by a sentence or two: A SNP’s association with a disease-causing mutation depends on how closely the polymorphism is (or polymorphisms are) to the actual mutation, while gene sequencing identifies the mutation itself. SNP testing is indirect; sequencing is direct.

The term “single nucleotide variant (SNV)” is easily confused with SNP and should EXPLICITLY be reserved only for mutations (variants) discovered by next generation sequencing (NGS). The authors use “genotyping” to cover both SNP detection and NGS detection, pointing out that the former “badly” genotypes rare mutations. Reader confusion would be lessened if genotyping was reserved only for the result of direct sequencing.

The crux of the central problem posed by SNP testing is explained in the first paragraph of the Discussion and in Figure 1. But it is not clear why very rare alleles will lie outside the main cluster. Neither the legends for ordinate and abscissa nor the title of the figure define. “strength” and “contrast.” "Automated clustering of dye signal intensities versus strength across multiple samples” is inadequate. No comment is made about why the center of the homozygous cluster for the graph for the more common variant has coordinates of x= -2, y=10.5; while the cluster in the graph for the rarer variant’s are x=1.5, y=10.5. What does the shift in the x value indicate? Presumably the homozygotes are normal, but that is never stated.

The words “call” or “calling” seems to me to be pure jargon and should either be defined or replaced with a more descriptive term. The last two sentences of the last paragraph of "Test methods" (lines 13-17) are particularly obtuse ("...a genotype missingness rate <5% and Hardy Weinberg P<1x10^{-6}, as is the first sentence under "Analyses:” “...with genomic positions present in the gVCF files and covered by >15 reads in the NGS data.”

The terms accuracy, reliability, and analytical validity are sometimes used synonymously, but in clinical chemistry they have different, distinct meanings. (See NIH-DOE Task Force on Genetic Testing, Chapter 3, (https://www.genome.gov/10002403/genetic-testing-reportchapter-3). I’d stick with analytical validity and remove the others.

In Figure 2 (a) and (b), the ranges in the legend overlap, e.g., if the upper range ends with 0.005% the next one should begin with 0.0051% or 0.006%.

Supplementary Figure 1 OR 2 should be included in the published paper, with a note that the algorithm for the other (e.g., the BiLEVE chip) was the same. Under “Patient and public involvement:” it is not quite right to say that “No patients or the public were involved in the design or implementation of this study…” as “Cancer registry data for breast, ovarian, prostate and pancreatic cancer was extracted for all participants. Logistic regressions were carried out to assess the relationship between test-positive participants and any BRCA-related cancer.” The authors must indicate that the analysis of these patients was conducted anonymously or with their consent if there were identifiers.

DTC SNP testing is used because it is cheaper and, perhaps, faster than NGS. How much cheaper and faster is it? Isn’t it likely that as NGS improves further that SNP testing in "medically-actionable" situations will become a relic? This paper should accelerate the transition.

Comments to the Editor

Men and women who have become physicians over the last twenty years may be aware of the terminology and methodologies used in this paper, and those trained in the last ten may even have performed mapping and sequencing experiments in college biology courses, but these groups are still a minority of physicians in practice today. To my mind, the findings in this paper, are sufficient for regulatory agencies to bar companies from offering DTC testing for clinical genetic testing for rare mutations. Short of that, when companies say “ask your doctor,” in advertising or providing results, doctors must be able to supply answers based on evidence. This paper could go a long way to educate them if it were to be written in easily understandable language without jargon or obscure terms. I’m confident this can be done, but it will require major revision. I am not asking that experiments be repeated or expanded, just that the paper can be understood by the vast majority of practicing physicians.
Comments:
Thank you for the opportunity to review the paper, "Very rare pathogenic genetic variants detected by SNP-chips are usually false positives: implications for direct-to-consumer genetic testing".

The objective of the paper—to determine the diagnostic accuracy of SNP-chips frequently used by direct-to-consumer (DTC) genetic testing companies for genotyping rare genetic variants—is an extremely important one given the explosive growth in the use of DTC testing by the public.

As the authors point out, the likelihood of receiving false results, particularly false positives, in the case of rare variants such as BRCA mutations are high and the implications of receiving these false results—either positive or negative—can be grave. People who receive false positives—something the authors demonstrate to be very common when using SNP-chips to detect rare variants such as BRCA mutations—are at risk of major emotional trauma as well as potentially unnecessary surveillance and life- and health-altering surgeries.

It is somewhat confusing that the authors do not at least touch on in this section the admittedly far rarer (because of the rarity of these variants), but still very problematic occurrence of receiving false negative results. While the NPV is 99.9%, the sensitivity for the combined SNP-chip testing is only 34.6%. So while most people receiving a negative result are in fact negative, there are clearly individuals who are positive for a pathogenic variant who are receiving false negative results with some regularity. People,
particularly women, who receive false negative results for pathogenic BRCA variants are at extremely high risk of developing cancer (breast/ovarian), a risk that could be greatly reduced through preventive surgeries and other interventions. I would recommend that the authors include a sentence on this fact (though I recognize they speak to it in general terms in the final section).

In terms of the accessibility of the paper, I think it could be improved. In particular, the use of the term “false discovery rate” was confusing, especially for those of us (many patients and clinicians) who were educated pre-1995. I hope the authors will work to ensure that this information is translated into an accessible format and made available to all those who would benefit from seeing and understanding it.

In short, this paper presents extremely important information that should be disseminated widely to clinicians, researchers, testing labs, and the public that the standard methodology used in DTC genetic testing is wholly unreliable for determining the presence of rare genetic variants and therefore should not be used to do so. Thank you for your efforts in this important area of research.

Additional Questions:
Please enter your name: Jill Holdren

Job Title: Patient/Research Advocate

Institution: The Light Collective

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: No

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests <A HREF='http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/declaration-competing-interests'target='_new'> (please see BMJ policy) </a>please declare them here:

Reviewer: 3

Comments:
Statistical Review
051528
This is a very interesting paper raising important issues, highlighting the harms which can be done from screening for rare conditions. However, it is currently a challenging read, but hopefully could be developed to more successfully highlight these important issues.
1. A key issue here is that the problem is driven largely (but not entirely) by the very low prevalence. Even the best tests have low positive predictive values when the pre-test probability of 0.00001. It would be helpful to make it clear in the paper that the problem is not specific to SNP-chips – they are an interesting example to focus the paper around, but it would be the same for other tests.

2. To the contrary, using many of these tests in people where the prevalence is higher (for example, using them for diagnosis when ‘people’ have become ‘patients’ as they present to health care with symptoms or family history) may possibly be useful, as the positive predictive value will be much higher.

3. Screening for rare diseases usually use a combination of two tests, for example an imaging test followed by a biopsy, as a way of resolving exactly this issue. The National Screening Committee undertakes serious consideration of test performance, and ensuring the likely benefits outweigh the potential harms before implementing any screening. It might help the authors to reflect on the issues that they consider, particularly when trying to draw out the lessons for clinical practice.

4. The terminology for measures of diagnostic accuracy that is used in the paper causes confusion, and the authors need to carefully review the terms that they are using. “False discovery rate” is not appropriate at all. “The likelihood of a true positive result” – it is unclear whether this is sensitivity, positive predictive value, or the proportion of true positives overall. Likewise “false positive rate” and “false negative rate” are ambiguous terms. Best to stick to sensitivity, specificity, positive and negative predictive values.

5. The terminology used to describe the tests, genetic investigations and conditions, whilst probably the daily language of the authors, is challenging for those of us who are not living their lives in a genetic world. This will be the case for many readers. We are used to thinking about index tests detecting particular target conditions. I have no idea of the significance of the conditions which are being identified – there is some global statement that they are clinically actionable, but please explain more, and please see whether the descriptions and technical language can be simplified. Also it seems that accuracy statistics are given for combinations of conditions, make sure this is properly explained.

6. There is no explanation given as to why NGS is suitable to be a reference standard. Please can you explain this? To be suitable there needs to be confidence that the findings from NGS are very likely to be correct. Please do not use the phrase “gold standard” – this is not suitable terminology for use in a test accuracy study.

7. Throughout the paper, results are presented without using confidence intervals. Whilst the overall sample size is large, actually some of the statistics quoted in the paper are based on very small subsets, and will have high uncertainty. It is essential that confidence intervals are presented with every statistic.

8. The low sensitivity of the tests is not well explained in discussion of the results (see 11 below). Sensitivities of 29.5% and 4.4% are quoted for ultrarare variants (these are the numbers where I expect CIs to be wide). This makes it clear that the PPV is low both because of rarity of the condition (which increases false positives), AND because it doesn’t do a good spotting the condition (reducing the true positives). The low sensitivity is a striking failure of the test, the low PPV is striking failure of putting a test in a pathway where the disease is ultrarare. Please make sure both aspects of this are reported. The discussion currently does not consider the generic problems of testing for rare conditions.

9. The conclusions starts using the phrase “analytical validity”. The study really is about “clinical validity” not analytical validity. Please check the terminology used.

10. There is a subtle issue here relating to how well a test performs in an individual compared to how well it works in a population. These issues are difficult to tease out. The text in the discussion talks about rare genetic variants being badly genotyped, whereas common genetic variants are well genotyped – to me this reads about how it is working within each individual. How well an individual in
genotyped for a genetic variant depends on the sensitivity and specificity of the test and is unrelated to disease prevalence. So how well an individual with variant is genotyped is described by the sensitivity, how well an individual without the variant is genotyped is described by the specificity. How well a test performs in a population moves terms of balancing risks of false positive against false negatives which depends in addition on the prevalence, and is where looking at the positive and negative predictive values.

11. Much of the early discussion concerns analytical problems with SNP testing for rare variants. It would be helpful to describe how these issues impact on sensitivity and specificity. Do they lead to variants being missed (low sensitivity) or do they wrongly test positive for a variant (low specificity).

12. The conclusion makes statements about the risk in patients with symptoms or family history. No data are presented in this paper to support these statements.

13. Can Table 1 please include the 2x2 data from which the accuracy statistics are calculated? This is required for the study to be included in future meta-analyses.

14. I am not sure that Figure 1 is helpful.

15. Figure 2 and Figure 3 might be better presented in a table. It is not clear what the error bars are here – putting the confidence intervals in a table would help.

Additional Questions:
Please enter your name: Jon Deeks

Job Title: Professor of Biostatistics

Institution: University of Birmingham

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: No

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests <A HREF='http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/declaration-competing-interests' target='_new'> (please see BMJ policy) </a> please declare them here: