PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form ([see an example](#)) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Developing A Brief Screening Instrument for Psychosocial Risk Associated with Genetic Testing – A Pan Canadian Cohort Study
AUTHORS	Esplen, Mary Jane; Cappelli, Mario; Wong, Jiahui; Bottorff, Joan; Hunter, Jon; Carroll, June; Dorval, Michel; Wilson, Brenda; Allanson, Judith; Semotiuk, Kara; Aronson, Melyssa; Bordeleau, Louise; Charlemagne, Nicole; Meschino, Wendy

VERSION 1 - REVIEW

REVIEWER	Angela Liegey Dougall, PhD Assistant Professor Department of Psychology The University of Texas at Arlington Arlington, Texas, USA I have no competing interests.
REVIEW RETURNED	07-Dec-2012

| THE STUDY | This manuscript describes the development of an instrument to timely identify individuals who are at risk for developing significant psychological distress following genetic testing for an Adult Onset Hereditary Disease. Development of a screening tool will assist genetic counselors in identifying patients who need help dealing with the stress associated with AOHD. As such, it is a topical area in the field of medicine.

First, I like the scale a lot. It conveniently captures important risk factors. I also think that the authors have approached the scale's development in an acceptable manner. I have some questions though concerning clarity of detail in the presentation of the methods and results and the method chosen for scoring the scale.

Below are detailed comments regarding The Study:

In the description of the selection of items in Study 1, it was indicated that genetic service providers rated the items on a 0 "excellently/definitely relevant" to 10 "very poor/definitely not relevant" scale. It is then stated that items were removed if they were rated below five. One of the numbers appears to be in error.

Under Materials in the Methods for Study 2, there needs to be more description of how the Hamilton Scales were (1) used in the analyses and (2) what they were measuring. For the first concern, please justify the cut-off scores that were used on the HAM-D and HAM-A. While the cut-off score of 12 on the HAM-D is indicative of mild depression, the cut-off score of 10 on the HAM-A falls below what is considered as mild anxiety symptoms. Please describe the relevance of these cut-offs to the population at hand. Additionally, |

	describe how these cut-off scores were used (e.g., If patients met either one or both cut-offs were they considered to meet caseness). Second, the Hamilton scales assess depression and anxiety but they are referred to in the text as distress. This is further confused by the authors referencing the results including the BSI and IES as depression and anxiety outcomes. There needs to be a clearer distinction among these outcome measures in order to strengthen the claims of convergent and predictive validity.
RESULTS & CONCLUSIONS	Under Participants in the Results section for Study 2, please describe whether the standardized telephone interview for the assessment of distress included the HAM-D and HAM-A as well as the IES and BSI or whether there were differences in collection. Additionally, detail what is meant by the phrase "up to 4 telephone calls were made." Clarify whether these calls were made just to reach participants or whether the standardized telephone interview took up to 4 calls to complete.

In this same section, describe which participants were missing follow-up data. Some of the missing data were not at random (participants had not received test results) and this pattern of missingness may have differed from a pattern in which participants had received test results but did not complete the study. Describe why a dichotomous variable (complete vs. incomplete) was more appropriate than breaking down the missing patterns (e.g., complete vs. did not receive test results vs. received test results but did not complete assessment). Referring back to my comment for the Methods, the presentation of the prevalence rates of the outcomes was confusing. There needed to be a clearer distinction among the outcome variables. The data related to the IES and the BSI were referred to as "depression and anxiety rate" in the text and "psychosocial well being" in the title for Table 2. Conversely, the HAM-D and HAM-A data were referred to in the text as "psychological distress." This problem continued throughout the Results.

In the Reliability and Factor Analysis section of the Results, the mean scores and standard deviations were given for the GPRI for the participants receiving testing for HD and cancer but not for hemochromatosis. Please describe if these patients were eliminated from analyses.

In the Validity section of the Results, please describe the characteristics of the patients who were correctly classified using the GPRI cut-off score (e.g., type of disease, particular items on the GPRI that were driving the classification, demographics, etc.). Whether or not there were discernible differences between patients who were and were not correctly classified will provide important information about the scale's usefulness.

My biggest concern is regarding the choice for scoring the GPRI. I like the content of the GPRI, but I do not believe that these risk factor items can be merely summed. There are three types of questions on this form. This is reinforced by the results of the factor analysis. The three factors corresponded to these three distinct types of risk factor questions (perceived impact, past mental health history, and personal experience with the disease). All three domains are important, but conceptually may represent independent risk factors. For example, it is well-known that prior history of psychiatric problems is a risk factor for development of psychiatric problems following a new index event (like genetic testing). |

Furthermore, each of these three types of risk factors may not be additive. Instead, there may be synergistic relationships, such as the combination of a positive mental health history and greater perceived impact or loss of a loved one and having a personal diagnosis. By taking a simple sum, all risk factor items are given an equal weight. I agree that the first two factors could be summed (separately), but I think the items in the third factor (personal disease experience) should be used individually. They can all be experienced independently. Instead of the ROC analyses (or in comparison to the ROC), I would like to see the authors use another type of predictive algorithm including the factor 1 score, factor 2 score, and the three items from factor 3 as predictors as well as possible interactions. I think the classification of participants may be improved. If not, it would further justify the simplistic approach. One of the strengths of the GPRI is that it can be completed quickly and scored immediately. If a more complicated algorithm is identified, a software application could be developed for easy use among genetic service providers. These providers are well-versed on similar algorithms with software applications (e.g., the Gail Model). Furthermore, the algorithm can be refined as more data become available from patients with diseases other than breast cancer or can be revised to include other relevant demographic and disease characteristics in the future.

The Discussion would benefit from a description of how the final items on the GPRI relate to the prior literature on risk factors for psychosocial impairment following genetic testing. An additional discussion on how interventions may be tailored to help decrease the risk incurred through the items in Factor 1 (perceived impact) would also be enlightening.

REVIEWER	Kurt D. Christensen, MPH, PhD Postdoctoral Fellow Division of Genetics, Department of Medicine Brigham and Women's Hospital and Harvard Medical School USA
REVIEW RETURNED	13-Dec-2012

| THE STUDY | Some analyses were confusing. I did not understand why the the BSI and IES, administered after testing, were not used to assess sensitivity/specificity of the GPRI. Also, the IES is anchored to a specific event, yet was administered prior to disclosure of test results. It is unclear what event is being assessed with the IES at baseline.

Also, wouldn't the specificity of the instrument be improved by incorporating the test result itself? The results certainly show a difference between participants testing positive and participants testing negative.

Anxiety and depression are different outcomes with different concerns. How/why are they lumped, and what do cutoffs mean? (i.e. 63 'cases' are identified. Does that mean 63 individuals scored above both HAM-D and HAM-A cutoffs? Just above cutoffs on one scale? What are those cutoffs?

Also, dropout is analyzed according to GPRI score, only the score hasn't been validated. The authors should report dropout by more |

	validated measures / demographic factors. All supplemental documents were relevant and did not raise concerns about the work.
RESULTS & CONCLUSIONS	I found the inconsistent definition of the sample universe confusing. The sample included in Table 1 (N=712) was not the same sample that was analyzed in the final sensitivity/specificity analyses (n=463). I would strongly recommend using the sample of completers throughout, although other presentation formats would be acceptable if a much greater description of the dropouts was provided. Again, I don't understand why the IES and BSI were not used to assess sensitivity and specificity, especially because cutoffs for concern based those two scales are clearly delineated. All comparison scales can be used for both construct validity and sensitivity/specificity analyses. Based on the numbers presented, I calculated that the instrument will have low positive predictive value (i.e., of anyone scoring above the author's proposed cutoff of 50, only 25% will actually be at distress levels of clinical concern. The authors should address this aspect. The authors are using scales with their own sensitivity/specificity problems rather than a DSM diagnosis of a mood disorder to assess the sensitivity and specificity of their proposed scale. What are the implications? Factor 3 of the proposed scale clearly lacks internal consistency (a Cronbach's alpha of .08 is very, very low). In fact the item, "I have/had a personal diagnosis" clearly does not belong with the other two items. I am sure the item is useful for predicting post-test mood disorders, but Factor 3 appears like it should be broken apart for the purposes of identifying subscales.
GENERAL COMMENTS	The authors have clearly invested much time an effort into developing an instrument that has the potential to help identify patients that may benefit from more rigorous follow-up. While I feel my critique is valid, I also feel it poorly reflects my opinion that the overall approach and work appear strong. There are a number of more minor issues I did not identify above, but should be mentioned: - Some minor aspects that affect the generalizability of findings are omitted, including patient race (psychometrics on these scales often vary by race) and education (relevant to interpreting the scale) - The first sentence of the abstract should finish the thought, "To develop a brief, reliable and valid screening instrument *to identify ___* for use...." The same comment applies to the introduction - Especially in the introduction, but throughout, there are minor style issues (e.g., run-on sentences, extra spaces, use of acronyms like MICRA that are not explained, "PubMed"). - In the methods section, it would be helpful to know who were "genetic service providers" (as in, "... further refined by genetic service providers"). Was that a specific team? How many? In the same paragraph, items were removed if they were "rated below

five." On average? By one person?

VERSION 1 – AUTHOR RESPONSE

the reviewers had the following main concerns:

1. Use of terms "Distress" vs. "Psychological Distress", and rationale or role of specific psychological scales

The general terms of "distress" and" psychological distress s" are frequently used interchangeably in the genetic testing literature and we had done the same. However, in our study we are measuring specific psychological symptoms and it is a helpful suggestion to make clearer to the reader what kind of distress we are referring to specifically. We have attempted to clarify where we are measuring specific psychological symptoms (e.g. depression or the specific anxiety associated with having a genetic test result). The revisions are highlighted in yellow in the text.

2. Further clarification of use of BSI and Hamilton Scale- to measure same or dissimilar symptoms?

Reviewers suggested further clarification on the rationale for the BSI and the use of the HAM-D and HAM-A during the various components of the study, given that they both measure symptoms of distress. Furthermore, it was unclear to the reviewers how we derived cut offs instituted, or their meaning in relation to their use in the current study and design of the instrument. We apologize if this was not clear in the first submission and have now addressed this concern.

In the original submission, the rationale that "the standardized interview based-rating scales should be used over subjective report scales as the principal outcome criterion in psychological distress both in general practice and in research trials [34]" and "Cases will be defined by established cut-offs for HAM-D >=12 [35] or HAM-A >=10 in the literature [36]" was located on page 14, in the results section. We realized that it should have appeared earlier in the instrument section and have now moved it up to Page-12 under outcome measures. We also added that these cut off points were established for populations in general practice, which was our study population. (These cut offs are not those recommended for psychiatric populations).

3. Scoring of the GPRI incorporating weights of risk factor clusters versus simple scoring approach.

Reviewers were concerned about the simple method of having a summation score and recommended an alternative scoring system with the use of "weights" that address the separate subscales. We had carefully considered alternatives of the scoring system in our analyses prior to the original manuscript submission and tested several approaches for predictive analysis. We aimed from the onset of the study to develop a user-friendly tool for genetics clinics to facilitate uptake of its use, while maintaining rigor and appropriate analyses.

However, these are excellent considerations to revisit. To address reviewers' concerns, here we present two possible alternative analyses: the binary logistic regression and the discriminant analysis. The binary logistic regression provides an odds ratio for each independent variable. We carried out binary logistic regression using one total-score, and repeated it with three factor sub-scores. The discriminant analysis provides the canonical discriminant function coefficient to indicate the "distance" between a new subject to the case vs. non-case, defined by the equation. Again, we carried out this analysis separately for one total-score versus three factor sub-scores. Results from these analyses are described below.

Binary logistic regression analysis – Single TOTAL score:

Variables in the Equation

B S.E. Wald df Sig. Exp(B)

Step 1a
totalscore .082 .012 45.790 1 .000 1.086
Constant -6.287 .715 77.364 1 .000 .002

95% C.I.for EXP(B)
Lower Upper
1.060 1.112

a. Variable(s) entered on step 1: totalscore.
The overall predicted probability is 86.6%

Binary logistic regression analysis –Subtotal for each of the THREE factors
Variables in the Equation

B S.E. Wald df Sig. Exp(B)

Step 1a
factor1total .083 .017 23.679 1 .000 1.086
factor2total .074 .024 9.258 1 .002 1.077
factor3total .108 .044 6.097 1 .014 1.114
Constant -6.454 .761 71.944 1 .000 .002

95% C.I.for EXP(B)
Lower Upper
1.051 1.123
1.027 1.129
1.023 1.214

a. Variable(s) entered on step 1: factor1total, factor2total, factor3total.
The overall predicted probability is 86.3%.

Discriminant analysis- single TOTAL score
Canonical Discriminant Function Coefficients
Function 1
totalscore .082
(Constant) -4.037

Using canonical discriminant function coefficient, D = -4.037 + 0.082*totalscore

The overall predicted probability is 86.6%

Discriminant analysis- Subtotal for each of the THREE factors
Canonical Discriminant Function Coefficients

Function 1
fac1totalscore .080
fac2totalscore .078
fac3totalscore .107
(Constant) -4.182

Using canonical discriminant function coefficient
D = -4.182 + 0.08*factor1total + 0.078*factor2total + 0.107*factor3total
The overall predicted probability is 86.3%.

Based on the above examinations, the overall predictive probability of cases could be improved slightly from 84% using ROC to 86% using equations with assigned coefficient to each factor score or to the total score. Based on these observations, we would like to suggest the use of ROC and a single total score. This approach reduced little of the predictive power yet it has the benefit of simplicity for clinicians to apply it in their primary care setting without having to use an additional scoring system.

4. Further clarification of characteristics of study drop outs or those who did not complete the follow-up measures to assess for potential differences or loss to follow up.

As per the reviewer's suggestion, we further separated the drop outs into those who did not receive genetic testing results, those who did not return the follow up questionnaires, and those who completed the follow up. Comparisons were made of these three groups and results are now reported on page 14.
" …Of the 712 participants, 85 (12%) did not receive genetic testing results at the scheduled follow up time and were not eligible for follow up measures on psychological symptoms in response to a genetic testing result. Of the remaining 627 participants, 152 (24%) did not return the self-administered follow up questionnaires and 12 (2%) submitted the follow up questionnaire package but did not complete a standardized telephone interview using HAM-D and HAM-A (up to 4 telephone calls were made to reach each participant for the interview). Therefore the final number of participants with complete follow-up data is 463 (463 over 627 eligible subjects, or 74%). The age, and baseline GPRI score between individuals who did not receive genetic testing results (age 51.4+12.7, GPRI 49.3+12.7), those who did not return the follow up questionnaires (age 48.1+11.6, GPRI 50.2+14.4) and those who completed follow up measures (age 50.1+12.8, GPRI 49.1+13.5) were compared. There was no statistically significant group differences (ANOVA and all post-hoc comparisons p>0.05)."

Reviewers were correct to suggest a separation between those who did not receive genetic testing results at the scheduled follow-up time versus those who did not return follow-up measures. By excluding those still waiting for genetic testing results from the eligible pool of follow-up participants, the actual response rate or participation rate should be 74% rather than 65% as was indicated in the initial submission.

5. In the Validity section of the Results, please describe the characteristics of the patients who were correctly classified using the GPRI cut-off score (e.g., type of disease, particular items on the GPRI that were driving the classification, demographics, etc.). Whether or not there were discernible differences between patients who were and were not correctly classified will provide important information about the scale's usefulness.

The reviewer raised a very important issue, that is, the characteristics of the "cases" who were classified using the GPRI cut-off score. We added a paragraph on page 17 to highlight this subgroup in terms of their characteristics, and if these were captured in the final GPRI tool.
"…Other demographic characteristics of these 63 subjects include: most were female and undergoing testing for BRCA1/2, which was similar to the whole sample of 712 (table 1). Compared with the

whole sample, these subjects had a slightly higher percentage of personal history of cancer (65% vs. 62%), higher rate of recent significant event of loss (56% vs. 47%), greater percentage reporting disease worries affecting mood (54.8% vs. 27%), having a feeling of sadness in the past month (46% vs. 17%) and anxiousness in the past month (33% vs. 17%). Our instrument captured all of these characteristics of this subsample."

6. Other more minor recommendations included: editing comments, further examples in discussion on how tool might guide intervention or follow up and limitations concerning sample.

We have provided where possible with our data set information on sample characteristics, identified study limitations and added some clarification or examples to our discussion section in relation to clinical use of the scale.

While completing the revisions addressing reviewers' concerns and incorporating their constructive suggestions, we realized that the added paragraphs increased the total word count of the paper to 5,115. We hope this would be acceptable, because these additional paragraphs, from our perspectives, greatly strengthened the manuscript and our study findings.

VERSION 2 – REVIEW

REVIEWER	Angela Liegey Dougall, PhD Assistant Professor The University of Texas Arlington
REVIEW RETURNED	29-Jan-2013

GENERAL COMMENTS	The authors have adequately addressed the previous concerns of the reviewers. This work provides a firm foundation on which future studies may further examine the usefulness of the screening instrument presented here.

REVIEWER	Kurt D. Christensen, MPH, PhD Postdoctoral Fellow Division of Genetics, Department of Medicine Brigham and Women's Hospital and Harvard Medical School USA
REVIEW RETURNED	23-Jan-2013

THE STUDY	Because there's no letter of response, it's hard for me to tell whether the authors ignored some comments for good reasons. Almost all the points below, I identified in my original review (I am omitting points that were unaddressed, but I feel are not essential to address). 1. In general, the Methods section is strong; but it is still unclear to me how the IES could be assessed at baseline before test results are disclosed and still be anchored to "the genetic test result." Do the authors meant that the IES was anchored to anticipated test results? 2. Given that the first specific predictor of distress mentioned in the Background section is, "While there is generally elevated distress ... among those who receive positive test results," it seems odd that test results are omitted from a scale predicting post-disclosure distress. If there's a good reason for omitting (not as predictive as

	items that were included, not really relevant for some reason), the authors should explain why.
	3. The Background section is not written with the same clarity as other sections.
RESULTS & CONCLUSIONS	While I feel the authors' methods were strong, again, they need to at least mention the issue that calculations of sensitivity and specificity are conducted not against a gold standard, but against scales that have their own sensitivity and specificity issues.
	I will just point out again that Factor 3 has incredibly poor internal consistency (Cronbach's alpha of .08). I looks like they're three uncorrelated items that were important predictors of future distress. I'd suggest describing them as such.

VERSION 2 – AUTHOR RESPONSE

One of the reviewers indicated that "the work is set upon a firm foundation which will facilitate future studies to examine the usefulness of the screening instrument". Below we respond to the first reviewer's comments and hope that these changes or responses will address the recommendation for minor revisions.

1. It is still unclear ..how the IES could be assessed at baseline before test results are disclosed and still be anchored to "the genetic test result". Do the authors mean that the IES was anchored to anticipated test results?

The reviewer is correct. This is a standard approach in the genetics literature. The IES is anchored, at baseline, to "anticipation of the genetic test result". And in follow up, as participants know their result, they respond in relation to their test result. We have tried to make this clearer in the paper and we believe readers will recognize this as a common approach when using the IES pre and post genetic testing.

2. It seems odd that test results are omitted from a scale predicting post-disclosure distress. …the authors should explain why.

The instrument designed for this study has a goal of being given prior to testing, and therefore, before any genetic test results become available. The items included in the instrument, such as perceptions of risk, personal history of mental health concerns, prior experiences with the disease, are robust items that are associated with post test psychological distress. They are selected from the literature and tested during our two phases of instrument development. These items are used to flag individuals who may develop psychological distress post genetic testing results. Because the instrument is designed to be used before individuals receive genetic testing results, the test result item is not included in the instrument

3. The Background section is not written with same clarity as other sections.

We have tried to edit and provide greater clarity in this section.

4. Further clarification of use of BSI and Hamilton Scale- to measure same or dissimilar symptoms?

This has been addressed in our earlier response. In page 12, we stated that "(Hamilton depression and anxiety) instruments were selected as main outcome measures based on the literature that the standardized interview based-rating scales should be used over subjective report scales as the

principal outcome criterion in psychological distress both in general practice and in research trials [34]."

5. Finally, other comments related to the issue of sensitivity and specificity being determined against self-report measures (e.g. IES, BSI) which may not be the gold standard as these scales may have their own sensitivity and specificity issues.

We recognize that the gold standard for measuring distress would be an in person interview, i.e., using the Structured Clinical Interview for DSM-IV (SCID). However, for this large study involving five centres across Canada, the time and cost required for SCID training and assessment prevents its implementation. We attempted to choose well-validated telephone interview measures, such as HAM-D and HAM-A, that have been used extensively in previous screening tool studies or in genetics and in other medical populations. Future studies could further validate the tool with either SCID or in person psychiatric interview among clinical genetics populations.

While HAM-D and HAM-A were used in our study for sensitivity, specificity measure, IES and BSI were used for construct validity. Both of them are self-report tools, as is GPRI, with acceptable psychometric properties, reasonable cost and relatively short time requirement to complete. This is highlighted on page 17.