PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Validation of breast cancer risk assessment tools on a French-Canadian population-based cohort
AUTHORS	Jantzen, Rodolphe; Payette, Yves; de Malliard, Thibault; Labbé, Catherine; Noisel, Nolwenn; Broët, Philippe

VERSION 1 – REVIEW

REVIEWER	Robert MacInnis Cancer Council Victoria, Australia
REVIEW RETURNED	24-Nov-2020

GENERAL COMMENTS	The authors of this manuscript performed a validation of 2 widely used breast cancer risk assessment tools: BCRAT and IBIS. Although these models have been validated in Canadian cohorts, they have not in French-Canadian cohorts. It is important to validate and compare the performance of these models in various cohorts. The results are based on only 131 incident breast cancer cases, and even less (58 cases) for validating PRS and combined scores. This should be mentioned as a limitation. If the study had just genotyped the remaining 73 incident cases, they could have turned this into a case-cohort study and achieved much more power for validating the PRS without having to genotype the rest of the cohort. This is assuming that those that were genotyped were a random sample of the clinic-based cohort, which it may not be given the authors state “selected to be genotyped through various scientific projects unrelated to breast cancer”. On the other hand, the characteristics in table S2 appear to be similar between the CC and CGC. I do not understand why the authors did not assess the IBIS model with PRS as the PRS is already fully integrated into their model (and thus the iCARE package is not needed to do this). The analysis treats incident deaths as censoring events. The BCRAT model computes absolute risk, while IBIS has an option to compute absolute risk, but the default is pure risk. It is unclear which version of the IBIS model was used. By treating death as censoring, if one were to compute a Kaplan-Meier curve to estimate cumulative risk of invasive breast cancer, one would be estimating the chance of getting breast cancer “if death from non-breast cancer could be eliminated.” Is this the risk one wants to use in counselling? It appears that key family history information was not available for the IBIS model – it is a little unclear what information was available and how it was inputted into the model. This may be an important limitation given family history and age at diagnoses is a key element of the model.

	The goodness of fit test for the 4 risk groups gave p-values < 0.05 for both BCRAT and IBIS. This should be stated in the abstract. Based on this, I would conclude that the models were not well-calibrated overall. This should be emphasised more, especially in the abstract. Stating that differences in results compared with other studies was due to a prediction horizon of 10 years seems a bit odd given the overall pattern is not likely to be markedly different between 5 and 10 years. While the rates between Canada and US are similar overall, there might be subtle differences for different age groups that might have an impact. I presume the UK rates were used for IBIS, but I couldn't find mention of this.

REVIEWER	Yurii Shvetsov University of Hawaii, USA
REVIEW RETURNED	13-Dec-2020

GENERAL COMMENTS	This manuscript describes a validation study of well-known breast cancer risk models in a population-based cohort in Canada. The paper is interesting and meaningfully contributes to the field. The results are nicely presented, in particular as plots of expected to observed incidence rate by strata of predicted risk. I have a number of comments and suggestions for the authors, outlined below. 1. The Introduction could be shortened, especially the description of the models being tested. Whereas a substantial part of the BCRAT paragraph deals with the model's application in Canada and thus is relevant, the following paragraph, describing the IBIS model, can be shortened. 2. Intro, p. 5, lines 45-50: it would be worthwhile to add that separate PRS have been constructed for specific races/ethnicities. 3. Methods, p. 6: It is unclear why two sources were used for breast cancer incidence information. Were they both incomplete? Did they offer different types of variables? (Some hint of that is given at the end of Discussion, but it's still unclear whether that was the only reason.) How was the information from the two sources reconciled? Were there any discrepancies for any study participants, and if so, which source was taken to be more reliable, and why? 4. Methods, p. 6, line 48: It is unclear what is meant by "women having an abnormal mammography". What was the time interval for abnormal mammography? Was it limited to the time before study recruitment, or during follow-up? How was the information on breast cancer diagnoses after abnormal mammography used? BCRAT includes breast biopsy (and atypical hyperplasia) variables, which is not the same as abnormal mammography. 5. Methods, p. 7, lines 6-9: Please explain why women from Phase 2 were not used. Was that because they did not have enough follow-up to assess 5-year observed breast cancer incidence? Also, it is stated that 12,062 women were genotyped, yet only 4,555 were included in the CGC. What is the reason that the rest of the 12,062 were not included? 6. Methods, pp. 7-8: Some of the terminology used in the paper appears to be non-standard. Please check the terms such as "hazard rates", "attributable hazard function", "risk score" and revise as necessary. For example, when the authors refer to a "risk score", do they mean relative risk estimates for the factors in the model? 7. Methods, section 2.4.2: It is unclear how using only 10% of the data for estimating risk factor distribution in the population would

	avoid optimism bias. Optimism bias typically refers to inflated performance measures when a model is built (relative risk estimates obtained) and tested using the same sample. In this case, the estimated population distribution of risk factors is used for constructing a baseline hazard function, per iCARE description. This baseline hazard does not depend on the levels of risk factors, i.e. it is the same for all participants, and thus it does not affect discriminatory performance of the model. Please explain or (preferably) use the entire sample to estimate the population distribution of risk factors. 8. Methods, section 2.4.3: It transpires that the authors have simply put together the relative risk estimates for all the factors from BCRAT and all SNPs from a particular PRS, and treated it as a "combined" model. I have a serious reservation about this approach. First, the PRS in question was obtained from a different population that the ones used for BCRAT construction. It is well known that geographically and racially different populations resulted in different published PRS. Thus, combining estimates from different populations is questionable. Second, the authors have not established that all SNPs from the PRS are orthogonal to all the factors in BCRAT (i.e. that the estimates would be independent). Without that (and I doubt that it is the case), one should enter all the factors into the same model and obtain new relative risk estimates that would account for confounding between the BCRAT and PRS factors. This was not done, thus the relative risk estimates in the "combined model" may be misleading. The authors should either remove this "combined model" or turn it into a teachable moment, state all the limitations outlined here, and conclude, based on the subpar calibration performance of this "combined model", that such mindless mechanical joining of disparate models is ill advised. 9. Methods, p. 9, lines 14-16: The statement about women with less than 5 years of follow-up contributing proportionately is unclear. If the authors refer to women who died during the course of follow-up, BCRAT accounts for the competing risk of death, so additional adjustment for that may not be necessary. 10. Methods, p. 9, line 29: Please state which goodness-of-fit test was used. Was it Hosmer-Lemeshow? 11. P. 9, line 39: Please insert "characteristic" after "Receiver operating". 12. Results: It is surprising that the c-statistic for IBIS differs so much between the CC and CGC cohorts: 63.42% vs. 59.63%, while that for BCRAT is very similar. Please comment on this in the Discussion.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Robert MacInnis

Institution and Country: Cancer Council Victoria, Australia

Please state any competing interests or state 'None declared': None declared

The authors of this manuscript performed a validation of 2 widely used breast cancer risk assessment tools: BCRAT and IBIS. Although these models have been validated in Canadian cohorts, they have not in French-Canadian cohorts. It is important to validate and compare the performance of these models in various cohorts.

The results are based on only 131 incident breast cancer cases, and even less (58 cases) for validating PRS and combined scores. This should be mentioned as a limitation. If the study had just genotyped the remaining 73 incident cases, they could have turned this into a case-cohort study and achieved much more power for validating the PRS without having to genotype the rest of the cohort. This is assuming that those that were genotyped were a random sample of the clinic-based cohort, which it may not be given the authors state "selected to be genotyped through various scientific projects unrelated to breast cancer". On the other hand, the characteristics in table S2 appear to be similar between the CC and CGC.

Answer: We completely agree with the reviewer that the power for validating the PRSs and combined scores would be improved with more genotyped incident cases. We have added this limitation in the Discussion section.

Unfortunately, we were unable to genotype more individuals in the CARTaGENE cohort for this study. Moreover, we confirm that participants were not randomly genotyped but for research projects unrelated to cancer, therefore genotyping only the 73 remaining incident cases would be problematic. We wanted to emphasize that this non-random selection was independent of breast cancer, which would have been an important selection bias in our study. However, as mentioned by the reviewer, the characteristics differences between the clinical-based cohort and the clinicogenetic-based cohort were modest, limiting the bias induced by the non-random genotyping.

I do not understand why the authors did not assess the IBIS model with PRS as the PRS is already fully integrated into their model (and thus the iCARE package is not needed to do this).

Answer: We thank the reviewer for its comment. As mentioned in the "2.4.3 Absolute risk using a combination of BCRAT and PRS" section, "*As the hazard function obtained from the IBIS model is not an output of the software, we cannot combine the IBIS and PRS information in this work.*". In this study, we used the iCARE package to update the BCRAT model based on the disease incidence rate and the distribution of risk factors in our population. Therefore, even if PRSs' relative risks are available, we need the relative risk of the IBIS model to add up their relative risks, which is not the case.

The analysis treats incident deaths as censoring events. The BCRAT model computes absolute risk, while IBIS has an option to compute absolute risk, but the default is pure risk. It is unclear which version of the IBIS model was used.

Answer: We thank the reviewer for its comment, and we agree that the version of the IBIS model used for the analysis was not stated clearly in our manuscript. In this work, we have taken into account the competing mortality in the IBIS program. Therefore, the IBIS absolute risk was computed, not the pure

risk. We have added this information in the first paragraph of the “2.4.1 Absolute risk using the BCRAT and the IBIS models” section.

By treating death as censoring, if one were to compute a Kaplan-Meier curve to estimate cumulative risk of invasive breast cancer, one would be estimating the chance of getting breast cancer “if death from non-breast cancer could be eliminated.” Is this the risk one wants to use in counselling?

Answer: Cause-specific hazard estimates are used when computing the absolute risk of breast cancer with BCRAT and IBIS. Thus, death from non-breast cancer is not eliminated but taken into account as a concurrent event in the absolute risk estimation, i.e., the expected proportion of cases.

The Kaplan-Meier estimator was used to compute the observed proportion of cases to take into account the follow-up time and to compare with the expected proportion of cases. We agree that this method, which is simple, can lead to biased estimates. However, in our study, we have very few women participants lost to follow-up, and the disease incidence is low.

It appears that key family history information was not available for the IBIS model – it is a little unclear what information was available and how it was inputted into the model. This may be an important limitation given family history and age at diagnoses is a key element of the model.

Answer: We agree with the reviewer that it is important to know how the variables and missing data were coded and that it was not stated clearly enough in the first version of the manuscript. We have added at the end of the “Variables extraction and coding” section in the Supplementary Methods a sentence on how the variables were coded: *How the variables were coded for the IBIS model can be found online (https://ems-trials.org/riskevaluator/), in the Documentation section, file “Risk program input file format (v6-8)”*.

Regarding the family history, only maternal and paternal history of breast cancer and maternal history of ovary cancer were available. We agree with the reviewer that this is a limitation of our study. Therefore, we have added this limitation in the Discussion. However, as the performance of the IBIS model remained good without the age at diagnoses and the other family history information, the IBIS model was penalized in our study and should be more accurate with more variables on family history.

The goodness of fit test for the 4 risk groups gave p-values < 0.05 for both BCRAT and IBIS. This should be stated in the abstract. Based on this, I would conclude that the models were not well-calibrated overall. This should be emphasised more, especially in the abstract.

Answer: We agree with the reviewer that we should have more clearly presented the results of the calibration in the abstract and that, even though BCRAT and IBIS have good mean calibration, it could be improved for risk subgroups.

We stated in the abstract (Results section) “BCRAT and IBIS had an overall expected-to-observed ratio of 1.01 [0.85-1.19] and 1.02 [0.86-1.21] **but with significant differences when partitioning by**

risk groups" and (Conclusion section) "BCRAT and IBIS **have good mean calibration that could be improved for risk subgroups,** and modest discriminatory accuracy". Moreover, in this revised version, we commented these results in more detail in the first and second paragraph of the Discussion section.

Stating that differences in results compared with other studies was due to a prediction horizon of 10 years seems a bit odd given the overall pattern is not likely to be markedly different between 5 and 10 years. While the rates between Canada and US are similar overall, there might be subtle differences for different age groups that might have an impact. I presume the UK rates were used for IBIS, but I couldn't find mention of this.

Answer: We agree with the reviewer that the differences obtained with a horizon of five or ten years should not be the main reason for explaining these differences as compared to the selection of the population and the year of the study. Thus, we removed this remark in the revised version.

Regarding the UK rates, we have added this information in the Discussion and in the "2.4.1 Absolute risk using the BCRAT and the IBIS models" sections.

Reviewer: 2

Reviewer Name: Yurii Shvetsov

Institution and Country: University of Hawaii, USA

Please state any competing interests or state 'None declared': None declared

This manuscript describes a validation study of well-known breast cancer risk models in a population-based cohort in Canada. The paper is interesting and meaningfully contributes to the field. The results are nicely presented, in particular as plots of expected to observed incidence rate by strata of predicted risk. I have a number of comments and suggestions for the authors, outlined below.

1. The Introduction could be shortened, especially the description of the models being tested. Whereas a substantial part of the BCRAT paragraph deals with the model's application in Canada and thus is relevant, the following paragraph, describing the IBIS model, can be shortened.

Answer: We have shortened the IBIS paragraph in the Introduction as suggested.

2. Intro, p. 5, lines 45-50: it would be worthwhile to add that separate PRS have been constructed for specific races/ethnicities.

Answer: We thank the reviewer for its comment. We have added this information in the Introduction of our manuscript with a reference as an example.

3. Methods, p. 6: It is unclear why two sources were used for breast cancer incidence information. Were they both incomplete? Did they offer different types of variables? (Some hint of that is given at the end of Discussion, but it's still unclear whether that was the only reason.) How was the information from the two sources reconciled? Were there any discrepancies for any study participants, and if so, which source was taken to be more reliable, and why?

Answer: We thank the reviewer for its comment and we agree that we should have provided more information about breast cancers in the manuscript. In practice, both sources may not be exhaustive: some women without a histologically confirmed cancers in the Breast Cancer Registry would only have information in the Quebec Health Insurance Board (RAMQ), and *vice versa*. To improve the clarity of the manuscript, we have added more information in the "2.1 Design and participants selection" section.

First we identified women with a histologically confirmed breast cancer in the Breast Cancer Registry. Then, for the women with an abnormal mammography without a histologically confirmed breast cancer, we identified breast cancers in the RAMQ database. Therefore, the Breast Cancer Registry was taken to be more reliable, while the RAMQ was used to identify women with abnormal mammography without a histologically confirmed breast cancer in the Breast Cancer Registry. This algorithm has a better predictive value than only using the Breast Cancer Registry or the RAMQ, as demonstrated in a previous published study[1], which was cited in our article (reference 28).

4. Methods, p. 6, line 48: It is unclear what is meant by "women having an abnormal mammography". What was the time interval for abnormal mammography? Was it limited to the time before study recruitment, or during follow-up? How was the information on breast cancer diagnoses after abnormal mammography used? BCRAT includes breast biopsy (and atypical hyperplasia) variables, which is not the same as abnormal mammography.

Answer: An abnormal mammography is a screening result coded as "abnormal" mammography in the Breast Cancer Registry. The two other possible responses are "normal" and "normal/benign lesion". Therefore, it is a lesion suspected of malignancy. We have clarified the definition of an abnormal mammography in the manuscript, section "2.1 Design and participants selection".

The abnormal mammography was only used in the algorithm to identify women with a breast cancer (see our previous response), while we used the breast biopsy results for BCRAT. Therefore, after identifying women with a breast cancer and the associated incidence date, we only included the women with an incidence date after the CARTaGENE inclusion date.

1 Théberge I, Institut national de santé publique du Québec, Direction systèmes de soins et services. Validation de stratégies pour obtenir le taux de détection du cancer, la valeur prédictive positive, la proportion des cancers in situ, la proportion des cancers infiltrants de petite taille et la proportion des cancers infiltrants sans envahissement ganglionnaire dans le cadre des données fournies par le programme québécois de dépistage du cancer du sein (PQDCS). Montréal: Direction des systèmes de soins et services, Institut national de santé publique; 2003.

5. Methods, p. 7, lines 6-9: Please explain why women from Phase 2 were not used. Was that because they did not have enough follow-up to assess 5-year observed breast cancer incidence? Also, it is stated that 12,062 women were genotyped, yet only 4,555 were included in the CGC. What is the reason that the rest of the 12,062 were not included?

Answer: We did not include the participants of the phase 2 as the family history of breast cancer was not available. We have clarified this point in the revised manuscript and added the sentence: "*as the family history of breast cancer was not available for the participants of the phase 2*" in the last paragraph of the section "2.1 Design and participants selection".

Moreover, the 12,062 individuals genotyped involved the whole CARTaGENE cohort, i.e., men and women, and participants of phases 1 and 2. The 4,555 individuals included only women and participants of the phase 1. To be clearer, we have modified the sentence in the "2.2 Genetic data" section: "Only a fraction of the **CARTaGENE** population cohort has been genotyped ~~**(n=12,062)**~~".

6. Methods, pp. 7-8: Some of the terminology used in the paper appears to be non-standard. Please check the terms such as "hazard rates", "attributable hazard function", "risk score" and revise as necessary. For example, when the authors refer to a "risk score", do they mean relative risk estimates for the factors in the model?

Answer: We thank the reviewer for its comment and agree that the use of these terms were confusing. We have replaced "hazard rates" with "hazard functions" and "attributable hazard function estimates" with "attributable risk". The "risk score" term was modified as it can be confusing, which sometimes referred to the relative risk.

7. Methods, section 2.4.2: It is unclear how using only 10% of the data for estimating risk factor distribution in the population would avoid optimism bias. Optimism bias typically refers to inflated performance measures when a model is built (relative risk estimates obtained) and tested using the same sample. In this case, the estimated population distribution of risk factors is used for constructing a baseline hazard function, per iCARE description. This baseline hazard does not depend on the levels of risk factors, i.e. it is the same for all participants, and thus it does not affect discriminatory performance of the model. Please explain or (preferably) use the entire sample to estimate the population distribution of risk factors.

Answer: We understand the reviewer's concern. We agree with the reviewer that the baseline hazard does not affect discriminatory performance of the model. However, it may affect the calibration of the model. As mentioned in the iCARE article[2]: "The risk factor distribution *F(Z)* plays a key role in calibrating the model to the marginal disease incidence rates in the underlying population. Thus, to carry out the calibration, the user must provide individual level data on the model risk factors for a sample that is representative of the underlying population.". Here, using a sample of 10% of the cohort provide information on the risk factor distribution but is not part of the validation. Thus, our

2 Choudhury PP, Maas P, Wilcox A, Wheeler W, Brook M, Check D, et al. iCARE: An R package to build, validate and apply absolute risk models. PLOS ONE. 2020 Feb 5;15(2):e0228198.

validation study uses only external information and avoids the criticism of overfitting our models to our dataset.

8. Methods, section 2.4.3: It transpires that the authors have simply put together the relative risk estimates for all the factors from BCRAT and all SNPs from a particular PRS, and treated it as a "combined" model. I have a serious reservation about this approach. First, the PRS in question was obtained from a different population that the ones used for BCRAT construction. It is well known that geographically and racially different populations resulted in different published PRS. Thus, combining estimates from different populations is questionable. Second, the authors have not established that all SNPs from the PRS are orthogonal to all the factors in BCRAT (i.e. that the estimates would be independent). Without that (and I doubt that it is the case), one should enter all the factors into the same model and obtain new relative risk estimates that would account for confounding between the BCRAT and PRS factors. This was not done, thus the relative risk estimates in the "combined model" may be misleading. The authors should either remove this "combined model" or turn it into a teachable moment, state all the limitations outlined here, and conclude, based on the subpar calibration performance of this "combined model", that such mindless mechanical joining of disparate models is ill advised.

Answer: We understand the reviewer's concern.

Regarding the first problem, as we conducted a validation study on a relatively homogeneous population, we have used PRSs based on the European population, which are the ancestors of most of the Quebecers, while the BCRAT model estimates are based on the (non-hispanic white) American women, which is close to the Canadian population.

Regarding the second problem, we agree with the reviewer that combining both clinical and genetic information in a simple additive way raises some concerns from an explanatory perspective, even though it may lead to good predictive performance. Indeed, if the genetic information is time-invariant, part of the clinical information is time-dependent (e.g., first live birth), with some information overlap between family history and PRS. However, since our interest in this work focuses on a predictive point of view, our oversimplified combination can be analyzed. However, as suggested by the reviewer, we should point out its explanatory limitations. Thus, we have added a sentence in the discussion regarding this problem: "It is worth noting that combining both clinical and genetic information in a simple oversimplified additive way has some limitations from an explanatory point of view even though it may lead to good predictive performance.".

9. Methods, p. 9, lines 14-16: The statement about women with less than 5 years of follow-up contributing proportionately is unclear. If the authors refer to women who died during the course of follow-up, BCRAT accounts for the competing risk of death, so additional adjustment for that may not be necessary.

Answer: We agree with the reviewer that the method used in this analysis can lead to biased estimates. We have redone the analyses without taking into account the follow-up time for the calibration analyses. The results were the same, with a very slight variation of the confidence interval that does not affect the results' interpretation, as very few women were lost to follow-up in our study. We updated the results in the Tables 1 and 2, and in the Results section. We also removed the

statement about the proportionately contribution of women with less than 5 years of follow-up in the 2.5.1 “Calibration” section.

10. Methods, p. 9, line 29: Please state which goodness-of-fit test was used. Was it Hosmer-Lemeshow?

Answer: We computed a global test statistic ($G = \sum_{i=1}^{4}(Oi - Ei)^2 / Ei$) and compared this latter to the critical value from the chi-squared distribution with four degrees of freedom. We have added this information in the manuscript, section “2.5.1 Calibration”. The statistic is different from the Hosmer-Lemeshow test since we work here with an independent validation sample.

11. P. 9, line 39: Please insert “characteristic” after “Receiver operating”.

Answer: We have added “characteristic” in the manuscript.

12. Results: It is surprising that the c-statistic for IBIS differs so much between the CC and CGC cohorts: 63.42% vs. 59.63%, while that for BCRAT is very similar. Please comment on this in the Discussion.

Answer: We agree with the reviewer that this result is surprising but might be linked to the smaller size of the clinicogenetic-based dataset as compared to the clinic-based dataset. We have added a sentence in the limitations paragraph of the Discussion section.

VERSION 2 – REVIEW

REVIEWER	Robert MacInnis Cancer Council Victoria, Australia
REVIEW RETURNED	04-Jan-2021

GENERAL COMMENTS	The manuscript is much improved and the authors have adequately addressed most of my comments. The only response I disagree with concerns assessing the IBIS mdoel with PRS. I do not see how the authors are able to assess IBIS without PRS but not IBIS with PRS. Both options provide outputs of absolute risk, which is all that is needed (you don't need the hazard function). I feel it would be a shame not to include it as it is clearly available and needs external assessment.

REVIEWER	Yurii Shvetsov University of Hawaii at Manoa, USA
REVIEW RETURNED	07-Jan-2021

GENERAL COMMENTS	The authors have adequately addressed most of the previous critiques, which has resulted in a much improved manuscript. However, a few outstanding minor issues from my original comments still remain, outlined below. 1. Regarding my original comment #4: The authors have not clarified the time period for which info on abnormal mammography was available and used. This is what I asked for, not for a definition of what abnormal mammography is. Please clarify the time period in the Methods. 2. Regarding my original comment #7: The authors have clarified that the 10% sample was used as required by iCARE to assess model calibration, not discrimination. However, the use of the term 'optimism bias' is confusing in this regard. Please delete the words 'To avoid the optimism bias' on p. 8, line 19. 3. Regarding my original comment #8: I appreciate that the authors now acknowledge the limitations of the 'oversimplified additive way', as requested in my comment. However, their statement 'even though it may lead to good predictive performance' is speculative and is not supported by the results of the paper, nor by any references. Please delete the words 'even though it may lead to good predictive performance' on p. 13, line 50.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Dr. Robert MacInnis, The University of Melbourne

Comments to the Author:

The manuscript is much improved and the authors have adequately addressed most of my comments. The only response I disagree with concerns assessing the IBIS mdoel with PRS. I do not see how the authors are able to assess IBIS without PRS but not IBIS with PRS. Both options provide outputs of absolute risk, which is all that is needed (you don't need the hazard function). I feel it would be a shame not to include it as it is clearly available and needs external assessment.

We agree with the reviewer that it would be interesting to assess IBIS with PRS, as BCRAT with PRS. In practice, we did not perform this analysis in the previous version of the manuscript as it relies on a different estimating strategy. For BCRAT+PRS, we have used the iCARE packages that provides an estimate of the baseline hazard rates taking into account the distribution of the risk score whereas this strategy is not possible for the IBIS model as the risk score is not an output of the package.

However and as underlined by the reviewer, the IBIS software allows for computing the absolute risk without or with PRS scores, but the cause-specific baseline hazard estimates are not the same of those derived from iCARE as in the last case (with PRS) they do not take into account the unknown distribution of PRS.

Bearing this in mind, we agree that computing the absolute risk from the IBIS software with the PRS is an option. Thus, as requested by the reviewer, we have used the IBIS breast cancer risk evaluation tool in adding the score for the PRS minus the logarithm of the expected value of the relative risk associated to the PRS in our population. This latter transformation is due to the fact that the baseline hazard rate can be approximated by the composite hazard divided by the expected value of the relative risk score in the underlying population. The results can be found in this table:

	BCRAT model / IBIS model	IBIS with Mavaddat	IBIS with Shieh	IBIS with Evans	IBIS with Wacholder
E/O	0.94 [0.73-1.22] 0.94 [0.73-1.22]	0.95 [0.73-1.22]	0.94 [0.73-1.22]	0.94 [0.73-1.22]	0.94 [0.73-1.22]
Goodness of fit	p=0.0415 p=0.268	p=0.470	p=0.519	p=0.993	p=0.627
Intercept	-2 [-4.4 - 0.2] -0.8 [-3.4 - 1.8]	-0.9 [-2.7 - 0.8]	-1.3 [-2.7 - 0]	-0.3 [-2.3 - 1.7]	-0.5 [-2.5 - 1.5]
Slope	0.5 [0 - 1] 0.8 [0.2 - 1.4]	0.8 [0.4 - 1.2]	0.7 [0.3 - 1]	0.9 [0.4 - 1.4]	0.9 [0.4 - 1.3]
C-index	59.13 [52.96-65.29] 59.63 [53.26-66]	62.73 [55.34-70.12]	63.83 [56.27-71.39]	63.35 [56.44-70.26]	64.21 [57.88-70.54]
C-indexes comparison with:					
BCRAT model	-	P=0.369	P=0.265	P=0.214	P=0.135
IBIS model	-	P=0.393	P=0.316	P=0.169	P=0.080
Sensitivity *	20.7% [11.2-33.4] 24.1% [13.9-37.2]	36.2% [24-49.9]	44.8% [31.7-58.5]	41.4% [28.6-55.1]	37.9% [25.5-51.6]
Specificity *	79% [77.7-80.3] 81.6% [80.4-	76.6% [75.2-77.9]	76.9% [75.5-78.2]	76.9% [75.6-78.2]	77.9% [76.6-79.2]

	82.8]				

The E/O were the same as the BCRAT and IBIS models. Intercepts and slopes were not different from zero and one, respectively. The c-indexes were all slightly higher than those obtained from the BCRAT and IBIS scores, but none of them were statistically different. Compared to the BCRAT and IBIS models, sensitivities values were higher while specificities values were lower.

Moreover, we also provided to the reviewer the results obtained without this transformation (shifted score) that shows bad calibration results:

	BCRAT model / IBIS model	**IBIS with Mavaddat**	**IBIS with Shieh**	**IBIS with Evans**	**IBIS with Wacholder**
Expected/Observed	0.94 [0.73-1.22] 0.94 [0.73-1.22]	1.77 [1.37-2.29]	4.25 [3.28-5.49]	1.54 [1.19-1.99]	1.39 [1.08-1.8]
Goodness of fit	p=0.0415 p=0.268	p<0.001	p<0.001	p=0.025	p=0.041
Intercept	-2 [-4.4 - 0.2] -0.8 [-3.4 - 1.8]	-1.4 [-2.9 - 0]	-2.4 [-3.3 - -1.6]	-0.8 [-2.6 - 1]	-0.8 [-2.7 - 0.9]
Slope	0.5 [0 - 1] 0.8 [0.2 - 1.4]	0.8 [0.3 - 1.2]	0.7 [0.3 - 1]	0.9 [0.4 - 1.4]	0.9 [0.4 - 1.3]
C-index	59.13 [52.96-65.29] 59.63 [53.26-66]	62.73 [55.34-70.12]	63.83 [56.26-71.39]	63.35 [56.44-70.26]	64.21 [57.88-70.55]
C-indexes comparison with:					
BCRAT model	-	p=0.369	p=0.265	p=0.214	p=0.135

IBIS model	-	p=0.393	p=0.317	p=0.169	p=0.080
Sensitivity *	20.7% [11.2-33.4]	81% [68.6-90.1]	96.6% [88.1-99.6]	77.6% [64.7-87.5]	69% [55.5-80.5]
	24.1% [13.9-37.2]				
Specificity *	79% [77.7-80.3]	35.4% [34-36.9]	8% [7.1-8.8]	40.5% [39-42.1]	49.4% [47.9-51]
	81.6% [80.4-82.8]				

We may hypothesize that these results are related to the estimation of the cause-specific baseline hazard rates that do not take into account the distribution of the PRS in our population. The differences in results observed between IBIS with PRS and BCRAT with PRS combined models reinforce this hypothesis and advocate for using this transformation.

In the revised version, we provided the results of the IBIS+PRS models (shifted score) and explained how it was computed.

In practice, we have added a section 2.4.4 “Absolute risk using a combination of IBIS and PRS” in the Methods section:

As the clinical risk score obtained from the IBIS model is not an output of the software, we cannot estimate the absolute risk associated with a combination of the IBIS clinical risk score and PRS using the iCARE package in the same way we did for BCRAT (see above). In practice, the version 8.0b of the IBIS risk evaluation tool allows to compute the absolute risk by incorporating the PRS scores, but these absolute risks are different from the ones that would be obtained with the iCARE package. Keeping in mind this issue, we have used the IBIS breast cancer risk evaluation tool and incorporate the PRS scores. More precisely, and for taking into account the distribution of the PRS, we incorporated a shifted PRS that corresponds to the PRS minus the logarithm of the expected value of the relative risk associated to the PRS in our population. This latter transformation is due to the fact that the baseline hazard rate can be approximated by the composite hazard divided by the expected value of the relative risk score in the underlying population ([34][3]).

We also have modified the Table 2 and the Figures 4, S1 and S2, and have added a paragraph in the Results section:

Regarding the IBIS + PRS combined models, the E/O were the same as the BCRAT and IBIS models (0.94 [0.73-1.22]) with non-significant goodness of fit tests (Table 2). All the combined models had an E/O that included one in each four risk groups (Figure 4). Intercepts and slopes were not different from zero and one, respectively (Table 2). The c-indexes were all slightly higher than those obtained from the BCRAT and IBIS scores, but none of them were statistically different. The discrimination for women at higher risk was also better for the Shieh and Mavaddat combined scores (down-left corner

3 Choudhury PP, Maas P, Wilcox A, Wheeler W, Brook M, Check D, et al. iCARE: An R package to build, validate and apply absolute risk models. PLOS ONE. 2020 Feb 5;15(2):e0228198.

of the ROC curves, Supplementary Figure S2). Compared to the BCRAT and IBIS models, sensitivities values were higher while specificities values were lower (Table 2).

Finally, we have added a paragraph in the Discussion section:

It should be noted that the combined IBIS+PRS models had a better calibration regarding the four risk groups compared to the BCRAT+PRS models. However, the absolute risk of IBIS combined models were not obtained with the same procedures as for BCRAT, which makes the results not straightforward to compare.

Reviewer: 2

Dr. Yurii Shvetsov, University of Hawai'i at Manoa

Comments to the Author:

The authors have adequately addressed most of the previous critiques, which has resulted in a much improved manuscript. However, a few outstanding minor issues from my original comments still remain, outlined below.

1. Regarding my original comment #4: The authors have not clarified the time period for which info on abnormal mammography was available and used. This is what I asked for, not for a definition of what abnormal mammography is. Please clarify the time period in the Methods.

We are sorry for not clarifying the reviewer's concerns. The time for which information on abnormal mammography was available was from May 15th, 1998 to December 31st, 2017 (i.e., the availability of the breast cancer registry). This information was available in the supplementary file, but we have added it in the “Design and participants selection” section of the manuscript.

2. Regarding my original comment #7: The authors have clarified that the 10% sample was used as required by iCARE to assess model calibration, not discrimination. However, the use of the term ‘optimism bias’ is confusing in this regard. Please delete the words ‘To avoid the optimism bias’ on p. 8, line 19.

As proposed by the reviewer, we have deleted the words “To avoid the optimism bias”.

3. Regarding my original comment #8: I appreciate that the authors now acknowledge the limitations of the ‘oversimplified additive way’, as requested in my comment. However, their statement ‘even though it may lead to good predictive performance’ is speculative and is not supported by the results of the paper, nor by any references. Please delete the words ‘even though it may lead to good predictive performance’ on p. 13, line 50.

As proposed by the reviewer, we have deleted the words “even though it may lead to good predictive performance”.

VERSION 3 – REVIEW

REVIEWER	Robert MacInnis Cancer Council Victoria, Australia
REVIEW RETURNED	09-Mar-2021

GENERAL COMMENTS	Thank you for presenting the extra IBIS+PRS analyses. I have no further comments.

REVIEWER	Yurii Shvetsov University of Hawaii at Manoa, USA
REVIEW RETURNED	24-Feb-2021

GENERAL COMMENTS	The authors have made the edits I had previously suggested, as well as addressed the other reviewer's critiques. In particular, the addition of IBIS + PRS model, requested by Reviewer 1, is welcome and makes for a stronger paper. I have no additional comments or critiques at this time.