Translate but validate: necessary steps in improving the use and utility of cancer risk models

M. B. Terry^{1,2}

Published online: 10 April 2020
© Springer Nature Switzerland AG 2020

Breast cancer is the leading cancer in women all over the world [1]. Given the challenges for development and validation of cancer risk models, it is not realistic to expect that separate risk models are developed and validated for each country. It is necessary to evaluate how existing models can be translated and applied in different settings. Translation, particularly into settings with differing incidence rates and risk factor patterns, requires validation. As politicians might advocate “trust but verify,” scientists should insist on “translate but validate.”

In this issue of *Cancer Causes and Control*, Pfeiffer and colleagues provide an exemplary study on how to apply “translate but validate” to two breast cancer risk models in the German population using both the German-EPIC cohort and the German Health Interview and Examination Survey (DEGS) [2]. To improve both the use and utility of risk models, it is important to consider key factors in the overall evaluation of these models which include the following: (1) independent, prospective validation; (2) comparison of multiple models; (3) assessment of the role of underlying incidence rates; (4) comparison of cohort versus population-based validation; and (5) model extensions and updates with new risk factors.

Independent, prospective evaluation

Given the relative rarity of cancers compared to other chronic diseases, it is challenging to develop new risk models and validate existing risk models. The prevalence of the outcome as well as the long induction times of most cancers, including breast cancer, makes development and validation reliant on large epidemiological studies. Case–control studies, efficient study designs for rare outcomes, can be used to estimate the discrimination of a model, that is, how likely a model can discriminate between individuals who are diagnosed with cancer versus those who are not. Commonly used discrimination measures include the area under the receiver operator curve (AUC) and the C-statistic. However, in addition to discrimination, model performance is also evaluated by calibration, that is, how often the model estimates match the actual observed cases.

To evaluate model calibration, prospective studies are essential as estimates from the risk model based on baseline epidemiological data need to be compared with the actual observed incidence that occurs in the cohort over time. Model calibration is typically evaluated based on the Hosmer–Lemeshow test. Calibration is often calculated overall, but caution should be exercised as good calibration on average may mask poor calibration for specific quantiles of risk. The only way to fully evaluate model performance based on both model discrimination and model calibration is through prospective studies. Evaluation and continued improvement of risk models requires out-of-forecast sampling, that is, validation in an independent sample. Again, given the relative rarity of cancer and the long induction time, such independent validation is often quite rare in contrast to risk models developed for more common diseases like heart disease.

✉ M. B. Terry
mt146@cumc.columbia.edu

¹ Department of Epidemiology, Mailman School of Public Health, Columbia University, 722 West 168th Street Room 1611, New York, NY 10032, USA

² Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY 10032, USA

Comparison of multiple models

Prospective, independent validation of multiple breast cancer risk models also provides perspective on how sensitive model performance estimates are to different assumptions guiding each risk model. For example, comparison of models that have very different model inclusion and modeling assumptions regarding the use of cancer family history, non-genetic data, and genetic data can provide valuable evidence on the quantitative impact of different model assumptions on the predicted absolute risk [3]. For example, comparing models that consider only first-degree family history data with other pedigree models can help evaluate the added value of second-degree family history. Likewise, comparison of models with similar non-genetic risk factors, but different assumptions about family history can shed light on the value of collecting information on ages of onset of cancers within pedigrees (for details see supplemental tables provided in [3]).

Pfeiffer and colleagues compare two breast cancer risk models: BCRAT [4] and BCRmod, a model which extends the framework of BCRAT to include modifiable factors [2]. A key difference between the two models is that BCRmod includes modifiable breast cancer risk factors (e.g., body mass index, hormone replacement therapy, and alcohol use). Both models include family history of breast cancer and reproductive breast cancer risk factors. Comparison across these models in both calibration and discrimination supports only minimal differences between these two models overall, but there are some differences when comparing performance in specific risk quantiles as well as selected subgroups based on age or risk factor category. The lack of major differences may also be explained, as the authors illustrate, by the lack of strong association with the additional modifiable factors included in BCRmod compared to BCRAT in the EPIC cohort. For example, some risk factors included in BCRmod like BMI were not associated with breast cancer risk in EPIC, while others like hormone replacement therapy were, but were missing to a large extent. Thus, caution may be warranted in concluding that the inclusion of these modifiable factors may not be meaningful in other settings. In addition, as BCRAT and BCRmod are not nested models, differences across the models cannot be used as a formal test whether the modifiable factors improve the model performance. A comparison of the two models based on common clinical cutoffs supports that more individuals increase in a category with BCRMOD than with BCRAT (Supplemental Table 4 in [2]).

Assessment of the role of underlying incidence rates

Translation of cancer risk models into other populations is most easily accomplished by changing the underlying cancer incidence rates to be specific to a given country. This means that the magnitude of the cancer risk factors is assumed to be similar across settings, which is not always a valid assumption. A major difference when applying German rates is seen in Fig. 1 in [2] where use of the German rates results in an underestimation of the upper quantiles in the BCRmod, whereas with the US rates there was an overestimation in the upper quantiles in both models. This difference may be clinically important as individuals in these upper quantiles may be more likely to be considered for more intensive screening and/or chemoprevention. Thus, using the German rates did have an impact on the conclusions about the model fit, particularly for BCRAT. The US breast cancer incidence rates are higher in every age group which likely reflects primarily increased frequency of mammography screening particularly in the older age groups and secondarily difference in risk factor patterns. As expected, use of the country-specific incidence rates largely impacted the calibration and not the discrimination. However, the fact that model performance was generally lower using German rates rather than US rates provides additional support for validation of risk models in new settings, rather than assuming that the models would perform better with country-specific rates.

Comparison of cohort versus population-based validation

In addition to validating the two models for 5 years and full study time period using the German-Epic cohort, Pfeiffer and colleagues report a 1-year calibration using population-based survey data from the German Health Interview and Examination Survey. In contrast to a general underestimation from both models when using EPIC, both models resulted in an overestimation when using the German Health Interview and Examination Survey in the older age groups. Although, both studies have population-based aspects to their sampling, a key difference between a cohort followed over 11 years and survey data followed for 1 year is the selection of who enters and stays in the cohort versus a broader snapshot of the population. In particular, the loss to follow-up over time may result in selection bias particularly if those who remain in the cohort differ with respect to underlying risk and/or mammography screening behavior, which the supplemental data supports. Even though risk

models calculate the expected based on the person time in the cohort and age at entry, they do assume that the loss to follow-up is non-differential. If women who join a cohort and/or stay in a cohort are different not only with respect to breast cancer risk factors, but also particularly with respect to mammography use, this could impact model evaluation. The approach taken by Pfeiffer and colleagues is a useful exercise particularly in knowing how to interpret validations in older women. The differences they observe support that broader caution should be made when comparing performance in cohorts that differ greatly in terms of underlying risk and risk factors and statistical approaches that weight results based on population data may be warranted.

Model extensions and updates with new risk factors

Both BCRM and BCRAT had low discrimination in the range of 60% in overall and for most subgroups. In general, the discrimination for most breast cancer risk models is only moderate; in the range of 60–68% with respect to discrimination with higher discrimination often seen when BRCA1/2 mutation data, pedigree data, and mammographic density data are available [3]. Typically, for clinical applications, including making recommendations regarding screening and chemoprevention, we would want a higher AUC [5]. Currently, much research is focused on artificial intelligence and other novel methods employing mammograms to improve absolute risk prediction [6]. Hopefully these methods will prove both efficient and valid, but still can only be used once a woman has had at least one mammogram which limit the applicability in settings where mammography is not available and in younger women where mammography is not sensitive. Other areas of exploration for extending risk models and improving discrimination include the use of polygenic risk scores and potential gene-environmental (GXE) interactions within risk models. To date, there has not been strong evidence of GXE, at least for the established breast cancer risk factors that have been tested in large consortia studies [7]. However, most risk models still do not include other risk factors like physical activity and aspirin which are consistent risk factors across the spectrum of absolute risk [8, 9].

A possible way forward

As breast cancer continues to increase around the world [1], we will need creative and robust approaches to translate and validate existing risk models in different settings. Breast cancer risk models have been developed primarily in studies with women of European ancestry. Translation to other settings requires knowing information

about relative risk factors and underlying incidence rates. Pfeiffer and colleagues [2] provide a rigorous framework to emulate in future validations including the need for (1) independent, prospective validation; (2) comparison of multiple models; (3) assessment of the role of underlying incidence rates; (4) comparison of cohort versus population-based validation; and (5) model extensions and updates with new risk factors. It is important to note, though, that following these steps even across populations that may be more similar to the populations where the models were developed suggests that understanding the drivers of heterogeneous model performance will be critical to the translation. Most revealing from their comparisons was the impact that changing country-specific incidence rates had on model performance. Their study reminds us that validation is essential as the assumption that performance would be better with German rates was not supported. In the end, better risk stratification will come from a better understanding of the key drivers of model performance which includes the need to continually examine the impact of underlying incidence rates and completeness and extension of breast cancer risk factors.

However, even without the fundamental challenges that “out-of-sample forecasts” pose to cancer risk models given the rarity of cancers and long induction time, it remains unrealistic to think that including more risk factors, unless that are of very high magnitude, will improve discrimination. Pooled studies will help improve precision, but often come at the expense of data harmonization which reduces risk factors to the “lowest common denominator” of categorization and often without repeated measures. This is illustrated by the missing data of key variables like hormone replacement therapy even within a single cohort like the German-EPIC cohort [2].

One alternative approach may be the one used in developing iPrevent [10] which is a validated switching algorithm which estimates absolute risk of breast cancer. Modifiable factors like tamoxifen use are applied outside of the model to counsel individual patients on the impact of initiating chemoprevention. In a similar way, other modifiable risk factors could be applied which would require only to have valid evidence from epidemiological studies for the relative risk estimate, but would not require incorporating the risk factor into an existing risk model. The advantage of this approach is not only that risk factor information would be valid and complete for the patient and provider, but that the risk assessment could be done at regular intervals and over time as a woman’s exposures change. In the end, we may need to consider different models based on whether mammography measures are available and in settings where mammography is not widely available evaluate whether models that utilize polygenic risk scores and non-genetic exposures to estimate absolute risk may be a potential approach in moving translation and validation forward.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 68(6):394–424. <https://doi.org/10.3322/caac.21492>
2. Husing A, Quante AS, Chang-Claude J, Aleksandrova Kaaks R, Pfeiffer R. (2020) Validation of two US breast cancer risk prediction models in German women. *Cancer Causes and Control* [in press].
3. Terry MB, Liao Y, Whittemore AS, Leoce N, Buchsbaum R, Zeinomar N, Dite GS, Chung WK, Knight JA, Southey MC, Milne RL, Goldgar D, Giles GG, McLachlan SA, Friedlander ML, Weideman PC, Glendon G, Nesci S, Andrulis IL, John EM, Phillips KA, Daly MB, Buys SS, Hopper JL, MacInnis RJ (2019) 10-year performance of four models of breast cancer risk: a validation study. *Lancet Oncol* 20(4):504–517. [https://doi.org/10.1016/S1470-2045\(18\)30902-1](https://doi.org/10.1016/S1470-2045(18)30902-1)
4. Gail M, Brinton L, Byar D et al (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 81:1879–1886
5. Wald NJ, Hackshaw AK, Frost CD (1999) When can a risk factor be used as a worthwhile screening test? *BMJ* 319(7224):1562–1565
6. Hopper JL, Nguyen TL, Schmidt DF, Makalic E, Song YM, Sung J, Dite GS, Dowty JG, Li S (2020) Going beyond conventional mammographic density to discover novel mammogram-based predictors of breast cancer risk. *J Clin Med* 26(9):E627. <https://doi.org/10.3390/jcm9030627>
7. Terry MB (2019) No multiplicative GXE interactions for breast cancer risk: have we reached a verdict or is the jury still out? *Int J Epidemiol*. <https://doi.org/10.1093/ije/dyz258>
8. Kehm RD, Genkinger JM, MacInnis RJ, John EM, Phillips KA, Dite GS, Milne RL, Zeinomar N, Liao Y, Knight JA, Southey MC, Chung WK, Giles GG, McLachlan SA, Whitaker KD, Friedlander ML, Weideman PC, Glendon G, Nesci S, kConFab investigators, Andrulis IL, Buys SS, Daly MB, Hopper JL, Terry MB (2020) Recreational physical activity and breast cancer risk: a cohort study of women selected for familial and genetic risk. *Cancer Res* 80(1):116–125. <https://doi.org/10.1158/0008-5472.CAN-19-1847>
9. Kehm RD, Hopper JL, John EM, Phillips KA, MacInnis RJ, Dite GS, Milne RL, Liao Y, Zeinomar N, Knight JA, Southey MC, Vahdat L, Kornhauser N, Cigler T, Chung WK, Giles GG, McLachlan SA, Friedlander ML, Weideman PC, Glendon G, Nesci S, kConFab investigators, Andrulis IL, Buys SS, Daly MB, Terry MB (2019) Regular use of aspirin and other non-steroidal anti-inflammatory drugs and breast cancer risk for women at familial or genetic risk: a cohort study. *Breast Cancer Res* 21:52. <https://doi.org/10.1186/s13058-019-1135-y>
10. Phillips KA, Liao Y, Milne RL, MacInnis RJ, Collins IM, Buchsbaum R, Weideman PC, Bickerstaffe A, Nesci S, Chung WK, Southey MC, Knight JA, Whittemore AS, Dite GS, Goldgar D, Giles GG, Glendon G, kConFab investigators, Cuzick J, Antoniou AC, Andrulis IL, John EM, Daly MB, Buys SS, Hopper JL, Terry MB (2019) Accuracy of estimates from the iPrevent breast cancer risk assessment and risk management tool. *JNCI Cancer Spectr* 3(4):pkz066. <https://doi.org/10.1093/jncics/pkz066>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.