Prediction of contralateral breast cancer: statistical aspects and prediction performance

Giardiello, D.

PREDICTION OF CONTRALATERAL BREAST CANCER

Statistical aspects and performance assessment

Daniele Giardiello

PREDICTION OF CONTRALATERAL BREAST CANCER
Statistical aspects and performance assessment
Daniele Giardiello

Prediction of contralateral breast cancer

Statistical aspects and performance assessment

Daniele Giardiello

Prediction of contralateral breast cancer

Statistical aspects and performance assessment

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 8 september 2022
klokke 13:45 uur

door

Daniele Giardiello
geboren te Cantù, Italië
in 1988

The work presented in this thesis was performed at the Netherlands Cancer Institute – Antoni van Leeuwenhoek, Amsterdam, the Netherlands, in cooperation with the Erasmus MC – University Medical Center Rotterdam, Rotterdam, the Netherlands and Leiden University Medical Center, Leiden, the Netherlands.

The research was funded by a grant from Alpe d'HuZes/Dutch Cancer Society (KWF Kankerbestrijding) under grant number A6C/6253.

Financial support for publication of this thesis was kindly provided by the Netherlands Cancer Institute – Antoni van Leeuwenhoek and the Netherlands Comprehensive Cancer Organization (IKNL).

Cover design: Maria Escala Garcia & Daniele Giardiello
Lay-out: Ilse Modder | www.ilsemodder.nl
Printing: Gildeprint Enschede | www.gildeprint.nl
ISBN: 978-94-6419-537-8

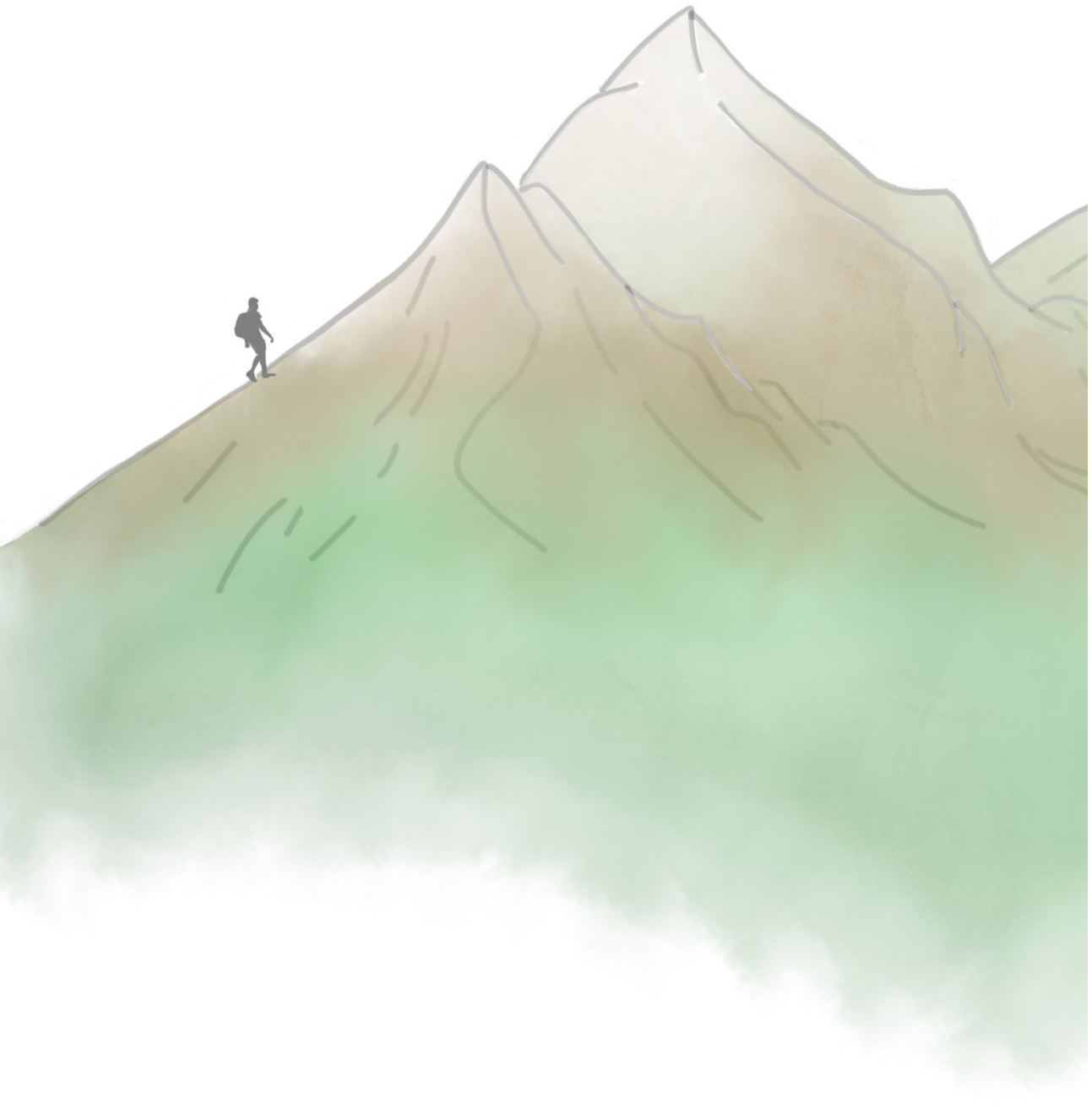
Promotores:
Prof. dr. M.K. Schmidt
Prof. dr. E.W. Steyerberg *Leiden University Medical Center*

Co-promotoren:
Prof. dr. M. Hauptmann *Brandenburg Medical School Theodor Fontane*

Leden promotiecommissie:
Prof. M.J.M. Broeders *Radboud UMC*
Prof. C.J. van Asperen
Prof. H. Putter
Prof. F.E. van Leeuwen *Vrije Universiteit Amsterdam*

To Salvatore Lo Vullo
(in memoriam, 22nd January 1971 - 2nd September 2021)

TABLE OF CONTENTS



Chapter 1

Introduction

INTRODUCTION

Breast cancer: developments and contemporary challenges

Breast cancer is the most common cancer in women in the world[1]. While the incidence of breast cancer has increased over the years, survival after breast cancer diagnosis has improved in the last 50 years due to earlier detection and advanced treatment modalities; for example, in the Netherlands, 10-year survival of first primary breast cancer approximately improved by almost 40% from 39% in 1961 to 76% in 2010[2,3]. As a consequence, breast cancer survivors may have substantial remaining lifetime to develop other cancers.

Breast cancer survivors are more likely to develop a new primary tumor in the opposite breast (defined as contralateral breast cancer) compared to healthy women to develop a first primary breast cancer[4-8]. Contralateral breast cancer is one of the biggest threats among breast cancer survivors: for example, in the Netherlands and in the United States contralateral breast cancer is the most common second primary cancer among women diagnosed with invasive breast cancer, accounting for 40-50% of all new second cancers[9,10]. About 5 out of 100 patients with invasive breast cancer will develop a contralateral breast cancer within 10 years since the diagnosis of the first primary breast cancer (10-year cumulative incidence between 4-7%, Figure 1)[11,12]. In addition, contralateral breast cancer patients have worse prognosis compared to patients with unilateral breast cancer[13-16]. Little is known about contralateral breast cancer among patients diagnosed with in situ breast cancer, a potential pre-invasive cancer occurring typically in the cell of milk ducts or lobules of the breast[17].

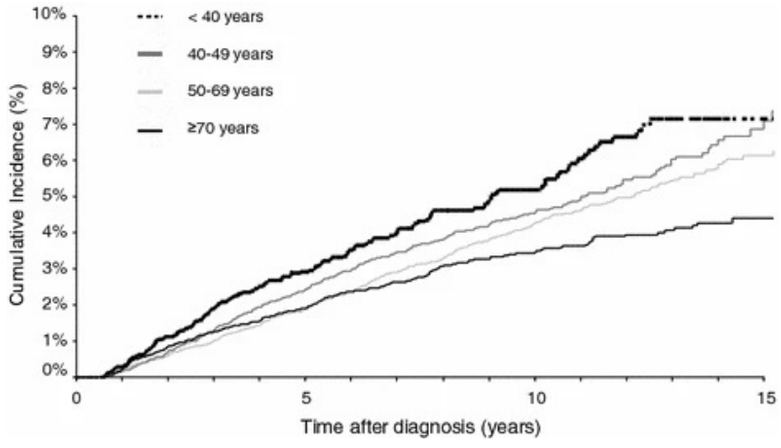


Figure 1: Cumulative incidence of contralateral breast cancer (CBC) by age and time since diagnosis (Data source: Netherlands Cancer Registry, N=45,229 index cancers; N=1,477 CBC[14], reproduced with permission of Breast Cancer Research and Treatment)

Contralateral breast cancer prevention: the preventive mastectomy

Although contralateral breast cancer risk is relatively low, an increasing number of patients with first breast cancer opt for a contralateral preventive mastectomy[18]. The rationale behind the contralateral preventive mastectomy is to avoid contralateral breast cancer with the consequent treatments and to potentially prevent death from a secondary primary breast cancer[19]. Among patients with highly elevated breast cancer risk contralateral preventive mastectomy is recommended, especially in patients with genetic predisposition or strong family history of breast cancer[20,21]. Currently, the contralateral preventive mastectomy is recommended in patients with germline mutations in *BRCA1* and *BRCA2* genes, although mutations in other breast cancer risk genes (*CHEK2, ATM* and *PALB2*) are suggested to also be associated with contralateral breast cancer risk[22-25].The 10-year contralateral breast cancer risk in women with *BRCA1*/2 germline mutations was estimated between 20-30% (Figure 2) [20-22,26]. Although the contralateral breast cancer risk is considerably high in women with *BRCA1/2* germline mutation, only between 1-5% of the European-descent general breast cancer population has a mutation in these genes[20,22,27]. Consequently, the choice of contralateral preventive mastectomy remains still debatable in a large part of general breast cancer population without any genetic predisposition. Furthermore, the increasing usage of (neo)adjuvant systemic therapies (chemo and endocrine therapy), intended to prevent breast cancer recurrences, has also been demonstrated to indirectly reduce the risk/incidence of contralateral breast cancer[12,18].

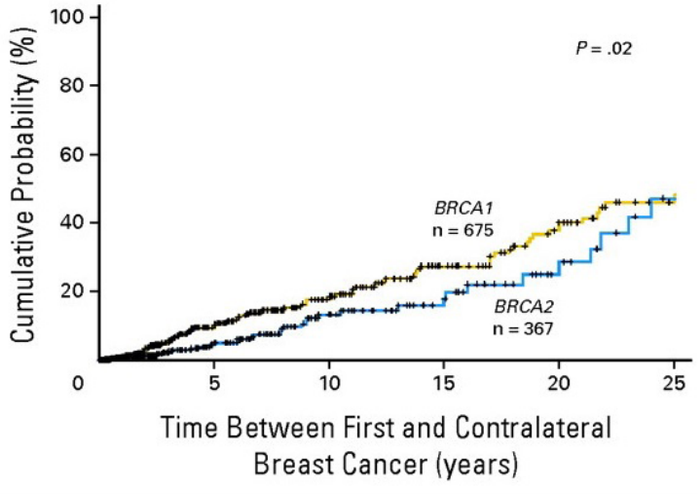


Figure 2: cumulative incidence of contralateral breast cancer in patients with *BRCA1/2* germline mutation (Data source: the German Consortium for Hereditary Breast and Ovarian Cancer, N=1,042; CBC=135[26], reproduced with permission from Journal of Clinical Oncology)

Decision-making for mastectomy

Psychological factors play a substantial role in a patient's decision regarding contralateral preventive mastectomy and in the outcome perceptions[18,28,29]. The risk reducing benefit of the surgery is commonly recognized. Contralateral preventive mastectomy mostly avoids a new breast cancer diagnosis and the subsequent new treatments[18,28]. As a consequence, contralateral preventive mastectomy might improve survival of first breast cancer patients. On the other hand, this intervention is not without negative consequences: approximately 30 out of 100 of women experience difficulties with body image, feminine identity, and sexual intimacy after surgery[18]. In addition, about 15-20% of patients who undergo contralateral preventive mastectomy may experience postoperative complications leading to higher medical costs and a potential lower quality of life[29-31]. Contralateral preventive mastectomy complications at (or about) the time of first breast cancer diagnosis may cause delays in initiating adjuvant systemic therapies increasing chance of breast cancer recurrences[29,30]. The benefit of avoiding the potential subsequent second breast cancer treatments should be correctly weighed against the costs of a negative body image and the potential postoperative complications. This harm and costs evaluation may be important during the consultation between patients and physicians to take a decision about the contralateral preventive mastectomy.

What is the role for prediction models in contralateral breast cancer prevention?

An appropriate risk prediction is crucial to more objectively quantify harms and benefit for a clinical decision making. Individualized contralateral risk prediction might help shared decision-making of physicians and patients about prevention strategies for those at high contralateral breast cancer risk, and to avoid unnecessary contralateral preventive mastectomies when contralateral breast cancer risk is low. Thus, contralateral breast cancer risk should be formally calculated to help patients and physicians during the decision making process towards preventive strategies, especially regarding contralateral preventive mastectomy[19]. This ambition may be successfully achieved under several conditions. First of all, the predicted risks should be sufficiently accurate and reliable. Good quality of data and performance assessment of the predicted risks is essential to evaluate prediction accuracy or to investigate the reasons causing inaccurate estimations. Secondly, the expected benefit of a clinical decision should be precisely quantified weighing pros and cons regarding preventive strategies. Last but not least, as long as predictions are accurate and the expected benefit is properly quantified, risk communication is really important in informing patients and physicians. Different methods of risk visualization may facilitate risk communication and the use of prediction tools in clinical practice[32].

Regression is the most widely statistical technique used to develop a risk prediction model and to provide absolute risk prediction[33]. A risk prediction model exploits the relation between predictors and the outcome of interest in a representative sample of patients. In

many cancer studies, the main outcome of interest is time until an event occurs, generally known as survival time. The methods considering survival time as outcome are defined as survival analysis [34]. However, when the event (e.g., death) does not occur in all individuals by the end of the follow-up, the true survival time is not known. This analytical problem is defined as censoring. Typically, censoring occurs when: a person does not experience the event of interest before the study ends, a person is lost to follow-up during the study, or a person withdraws from the study. The latter may happen when a competing event occurs. Let contralateral breast cancer event be the event of interest, patients may withdraw from the study because of dying. Therefore, death is a competing event that precludes the contralateral breast cancer from happening. Statistical models accounting for both censoring, and competing risks exist and are widely used in clinical practice such as in cardiology and oncology[35-37]. The most common statistical regression models for survival analysis with or without competing risks are the Cox proportional hazard regression and the Fine and Gray regression model[37].

Evaluation of performance and utility of risk prediction models

The statistical performance of a risk prediction model is important to validly support decision-making. A risk prediction model's performance is measured usually in terms of discrimination and calibration. The former is the ability of the model to identify subjects with good outcome and with poor outcome[38,39]. The latter is the agreement between observed and predicted outcome[38]. Both can be evaluated in the development data as internal validation or in independent data as external validation. The latter assessment provides an indication about the generalizability and the transportability of the risk prediction model in a new setting[40,41]. An increasing number of performance measures have been proposed in the last two decades for survival models. Many extend to the case of competing risks. However, clear guidance is lacking for a comprehensive assessment of the performance for survival and competing risks models [42-44]. Moreover, a model may show good performance in terms of discrimination and calibration, while both measures are unable to provide an answer to the question whether a risk prediction model should be used in practice to guide clinical decision making[45].

Net benefit is a relative novel measure to evaluate the clinical utility of risk prediction models and diagnostic tests weighting benefit and harms of a clinical decision making in public health[46,47]. Early detection and disease prevention are two of the most important goals in medicine. Physicians generally accept to recommend to persons or patients a certain number of unnecessary preventive strategies or treatments for the benefit to early detect or prevent a disease. This implies that the cost of missing the early detection or prevention of a certain disease (defined as false negatives) is typically more important than the cost of unnecessary preventive strategies or treatments recommendation (defined as false positives). The nationwide mammography screening program is a clear

example: public health physicians accept to recommend a high number of unnecessary screenings in a large population of healthy women to early detect a breast cancer with the aim to anticipate treatment and improve prognosis. This is also feasible because mammography screenings are considered as an acceptable safe procedure with a low number of side effects. In disease prevention, treatments or preventive strategies like surgeries may be more harmful than the mammography screening. For example, as previously mentioned, a certain number of complications after a contralateral preventive mastectomy may be possible.

Imagine having to decide about contralateral preventive mastectomy in 1000 patients diagnosed with first breast cancer. As previously reported, about 50 out of 1000 patients will develop a cancer in the contralateral breast in 10 years (i.e., as previously reported an expected 5% 10-year cumulative incidence). Suppose physicians recommend contralateral preventive mastectomy to all breast cancer patients irrespective of their age, potential germline mutations, family history and the other first breast cancer characteristics. We define this strategy as "intervention to all". This means we prevent 50 contralateral breast cancers (i.e., 5% benefit) at the cost of 950 unnecessary surgeries (i.e., 95% harm). Now, suppose that another strategy (named "alternative strategy", e.g., using a risk prediction model) is available and it reduces the number of unnecessary surgeries to 450, but preventing only 40 contralateral breast cancers. Is reducing 500 unnecessary surgeries at the cost of not preventing 10 contralateral breast cancers a good trade-off? To answer this question, it is important that physicians define how many patients should unnecessary undergo the contralateral preventive mastectomy to prevent one contralateral breast cancer. For example, a physician thinks that no more than 25 patients should undergo the surgery to prevent one contralateral breast cancer: this implies that not preventing a contralateral breast cancer is twenty-four times more harmful than undergoing an unnecessary preventive surgery. This 1:25 ratio would imply a decision threshold of 4%, and a relative weight of unnecessary surgery as 1/24 that of missing one contralateral breast cancer. The result of the net benefit calculation is reported in Table 1.

The net benefit is 1% and 2% for the strategy "interventions to all" and "alternative strategy", respectively. In other words, assuming the same number of unnecessary interventions, the "alternative strategy" prevents 20 contralateral breast cancers per 1000 patients at risk. The "intervention to all" strategy leads to prevent less contralateral breast cancers (i.e., 10) than the "alternative strategy". Thus, the net benefit of the "alternative strategy" is higher than the "intervention to all" strategy using 1/24 as weight of benefit and harms. However, physicians may have different opinions about how many patients should undergo unnecessary surgeries to prevent one disease. For this reason, net benefit calculation may be possible to define which strategy has higher net benefit using different exchange rates[46-48].

Table 1: an example of net benefit calculation

How many unnecessary mastectomies a physician is willing to accept to prevent a contralateral breast cancer?	Number of patients with first breast cancer: 1000 Expected number of contralateral breast cancers in 10 years: 50				
	Exchange rate	Strategies	Benefit	Harm	Net Benefit = benefit - (harm × exchange rate)
25	1/24	Interventions to all	5% (50/1000)	95% (950/1000)	5% - (95% ×1/24) = 1%
		Alternative strategy	4% (40/1000)	45% (450/1000)	4% - (45% ×1/24) = 2%

The goal of this thesis is to develop, validate and evaluate the potential clinical utility of a contralateral breast cancer prediction model to provide contralateral breast cancer prediction at 5 and 10 years since diagnosis of first primary breast cancer. Frameworks of assessing prediction performance in time-to-event models with or without competing risks are proposed using motivating examples in breast cancer.

THESIS OUTLINE

We set out to develop a risk prediction model for contralateral breast cancer. named PredictCBC, using international population-based and hospital-based studies (Table 2, Chapter 2).

Table 2: Data sources used in the thesis

Source	Country	Description	Chapter
Amsterdam Breast Cancer Study (ABCS)	the Netherlands	Hospital-based study	2-3-4
Breast Cancer Association Consortium (BCAC)	International	Population and hospital-based studies	2-3-4
Breast Cancer Outcome Study of Mutation (BOSOM)	the Netherlands	Hospital-based study	2-3-4
Erasmus Medical Center (EMC)	the Netherlands	Hospital-based study	2-3-4
Hereditary Breast and Ovarian cancer study (HEBON)	the Netherlands	Population-based study*	4
the Netherlands Cancer Registry (NCR)	the Netherlands	Population-based study	2-3-4-5-7

*selection through clinical genetic centers

Using the same data (Table 2), we evaluated and compared the prediction performance of PredictCBC with other tools available to calculate the contralateral breast cancer risk in clinical practice: CBCrisk and the Manchester formula (Chapter 3). We updated PredictCBC models using more clinical and genetic information available including body mass index, parity, *CHEK2* c.1100del, and polygenic risk score to potentially improve the contralateral breast cancer risk prediction performance for decision making (Chapter 4). We estimated the contralateral breast cancer risk in patients with ductal carcinoma in situ, a possible precursor of breast cancer since less is known about contralateral breast cancer risk in comparison with invasive breast cancer patients (Chapter 5). Finally, we provide frameworks of how to assess prediction performance in time-to-event models with and without competing risks using motivating examples in breast cancer as a guidance for researchers and practitioners interested in risk prediction (Chapter 6 and Chapter 7).

REFERENCES

1 Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424, doi:10.3322/caac.21492 (2018).

2 Netherlands Cancer Registry (NCR). *Survival and prevalence of cancer*, <https://www.cijfersoverkanker.nl> (2016).

3 van der Meer, D. J. *et al.* Comprehensive trends in incidence, treatment, survival and mortality of first primary invasive breast cancer stratified by age, stage and receptor subtype in the Netherlands between 1989 and 2017. *Int J Cancer* **148**, 2289-2303, doi:10.1002/ijc.33417 (2021).

4 Brenner, D. J. Contralateral second breast cancers: prediction and prevention. *J Natl Cancer Inst* **102**, 444-445, doi:10.1093/jnci/djq058 (2010).

5 Mariani, L. *et al.* Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Res Treat* **44**, 167-178, doi:10.1023/a:1005765403093 (1997).

6 Veronesi, U. *et al.* Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *N Engl J Med* **347**, 1227-1232, doi:10.1056/NEJMoa020989 (2002).

7 Schaapveld, M. *et al.* Risk of new primary nonbreast cancers after breast cancer treatment: a Dutch population-based study. *J Clin Oncol* **26**, 1239-1246, doi:10.1200/JCO.2007.11.9081 (2008).

8 Cheung, K. J. & Davidson, N. E. Double Trouble: Contralateral Breast Cancer Risk Management in the Modern Era. *J Natl Cancer Inst* **111**, 641-643, doi:10.1093/jnci/djy203 (2019).

9 Curtis, R. E., Ron, E., Hankey, B. F. & Hoover, R. N. in *New malignancies among cancer survivors: SEER Cancer Registries, 1973–2000* (ed National Institutes of Health (NIH)) pp 185-2005 (2006).

10 Soerjomataram, I. *et al.* Risks of second primary breast and urogenital cancer following female breast cancer in the south of The Netherlands, 1972-2001. *Eur J Cancer* **41**, 2331-2337, doi:10.1016/j.ejca.2005.01.029 (2005).

11 Chen, Y., Thompson, W., Semenciw, R. & Mao, Y. Epidemiology of contralateral breast cancer. *Cancer Epidemiol Biomarkers Prev* **8**, 855-861 (1999).

12 Kramer, I. *et al.* The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype. *J Natl Cancer Inst*, doi:10.1093/jnci/djz010 (2019).

13 Vichapat, V. *et al.* Prognosis of metachronous contralateral breast cancer: importance of stage, age and interval time between the two diagnoses. *Breast Cancer Res Treat* **130**, 609-618, doi:10.1007/s10549-011-1618-8 (2011).

14 Schaapveld, M. *et al.* The impact of adjuvant therapy on contralateral breast cancer risk and the prognostic significance of contralateral breast cancer: a population based study in the Netherlands. *Breast Cancer Res Treat* **110**, 189-197, doi:10.1007/s10549-007-9709-2 (2008).

15 Langballe, R. *et al.* Mortality after contralateral breast cancer in Denmark. *Breast Cancer Res Treat* **171**, 489-499, doi:10.1007/s10549-018-4846-3 (2018).

16 Hartman, M. *et al.* Incidence and prognosis of synchronous and metachronous bilateral breast cancer. *J Clin*

Oncol **25**, 4210-4216, doi:10.1200/JCO.2006.10.5056 (2007).

17 Mariotti, C. Ductal Carcinoma in Situ of the Breast. *Springer International Publishing* (2018).

18 Murphy, J. A., Milner, T. D. & O'Donoghue, J. M. Contralateral risk-reducing mastectomy in sporadic breast cancer. *Lancet Oncol* **14**, e262-269, doi:10.1016/S1470-2045(13)70047-0 (2013).

19 Narod, S. A. Bilateral breast cancers. *Nat Rev Clin Oncol* **11**, 157-166, doi:10.1038/nrclinonc.2014.3 (2014).

20 van den Broek, A. J. *et al.* Impact of Age at Primary Breast Cancer on Contralateral Breast Cancer Risk in BRCA1/2 Mutation Carriers. *J Clin Oncol* **34**, 409-418, doi:10.1200/JCO.2015.62.3942 (2016).

21 Kuchenbaecker, K. B. *et al.* Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317**, 2402-2416, doi:10.1001/jama.2017.7112 (2017).

22 Weischer, M. *et al.* CHEK2*1100delC heterozygosity in women with breast cancer associated with early death, breast cancer-specific death, and increased risk of a second breast cancer. *J Clin Oncol* **30**, 4308-4316, doi:10.1200/JCO.2012.42.7336 (2012).

23 Broeks, A. *et al.* Excess risk for contralateral breast cancer in CHEK2*1100delC germline mutation carriers. *Breast Cancer Res Treat* **83**, 91-93, doi:10.1023/B:BREA.0000010697.49896.03 (2004).

24 Robson, M. E. *et al.* Association of Common Genetic Variants With Contralateral Breast Cancer Risk in the WECARE Study. *J Natl Cancer Inst* **109**, doi:10.1093/jnci/djx051 (2017).

25 Fanale, D. *et al.* Detection of Germline Mutations in a Cohort of 139 Patients with Bilateral Breast Cancer by Multi-Gene Panel Testing: Impact of Pathogenic Variants in Other Genes beyond BRCA1/2. *Cancers (Basel)* **12**, doi:10.3390/cancers12092415 (2020).

26 Graeser, M. K. *et al.* Contralateral breast cancer risk in BRCA1 and BRCA2 mutation carriers. *J Clin Oncol* **27**, 5887-5892, doi:10.1200/JCO.2008.19.9430 (2009).

27 Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *J Mammary Gland Biol Neoplasia* **9**, 221-236, doi:10.1023/B:JOMG.0000048770.90334.3b (2004).

28 Ager, B. *et al.* Contralateral prophylactic mastectomy (CPM): A systematic review of patient reported factors and psychological predictors influencing choice and satisfaction. *Breast* **28**, 107-120, doi:10.1016/j.breast.2016.04.005 (2016).

29 Agarwal, S., Pappas, L., Matsen, C. B. & Agarwal, J. P. Second primary breast cancer after unilateral mastectomy alone or with contralateral prophylactic mastectomy. *Cancer Med*, doi:10.1002/cam4.3394 (2020).

30 Keskey, R. C. *et al.* Cost-effectiveness Analysis of Contralateral Prophylactic Mastectomy Compared to Unilateral Mastectomy with Routine Surveillance for Unilateral, Sporadic Breast Cancer. *Ann Surg Oncol* **24**, 3903-3910, doi:10.1245/s10434-017-6094-x (2017).

31 Parker, P. A. *et al.* Prospective Study of Psychosocial Outcomes of Having Contralateral Prophylactic Mastectomy Among Women With Nonhereditary Breast Cancer. *J Clin Oncol* **36**, 2630-2638, doi:10.1200/JCO.2018.78.6442 (2018).

32 Van Belle, V. & Van Calster, B. Visualizing Risk Prediction Models. *PLoS One* **10**, e0132614, doi:10.1371/journal.pone.0132614 (2015).

33 Harrell, F. E., Jr. Regression Modeling Strategies with applications to linear models, logistic and ordinal regression, and survival analysis. *Springer Series in Statistics* **2nd edition** (2015).

34 Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival analysis part I: basic concepts and first analyses. *Br J Cancer* **89**, 232-238, doi:10.1038/sj.bjc.6601118 (2003).

35 Haller, B., Schmidt, G. & Ulm, K. Applying competing risks regression models: an overview. *Lifetime Data Anal* **19**, 33-58, doi:10.1007/s10985-012-9230-8 (2013).

36 Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation* **133**, 601-609, doi:10.1161/CIRCULATIONAHA.115.017719 (2016).

37 Wolbers, M., Koller, M. T., Witteman, J. C. & Steyerberg, E. W. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* **20**, 555-561, doi:10.1097/EDE.0b013e3181a39056 (2009).

38 Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128-138, doi:10.1097/EDE.0b013e3181c30fb2 (2010).

39 Pencina, M. J. & D'Agostino, R. B., Sr. Evaluating Discrimination of Risk Prediction Models: The C Statistic. *JAMA* **314**, 1063-1064, doi:10.1001/jama.2015.11082 (2015).

40 Steyerberg, E. W. & Harrell, F. E., Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* **69**, 245-247, doi:10.1016/j.jclinepi.2015.04.005 (2016).

41 Austin, P. C. *et al.* Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol* **79**, 76-85, doi:10.1016/j.jclinepi.2016.05.007 (2016).

42 van Houwelingen, H. C. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* **19**, 3401-3415 (2000).

43 Kamarudin, A. N., Cox, T. & Kolamunnage-Dona, R. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med Res Methodol* **17**, 53, doi:10.1186/s12874-017-0332-6 (2017).

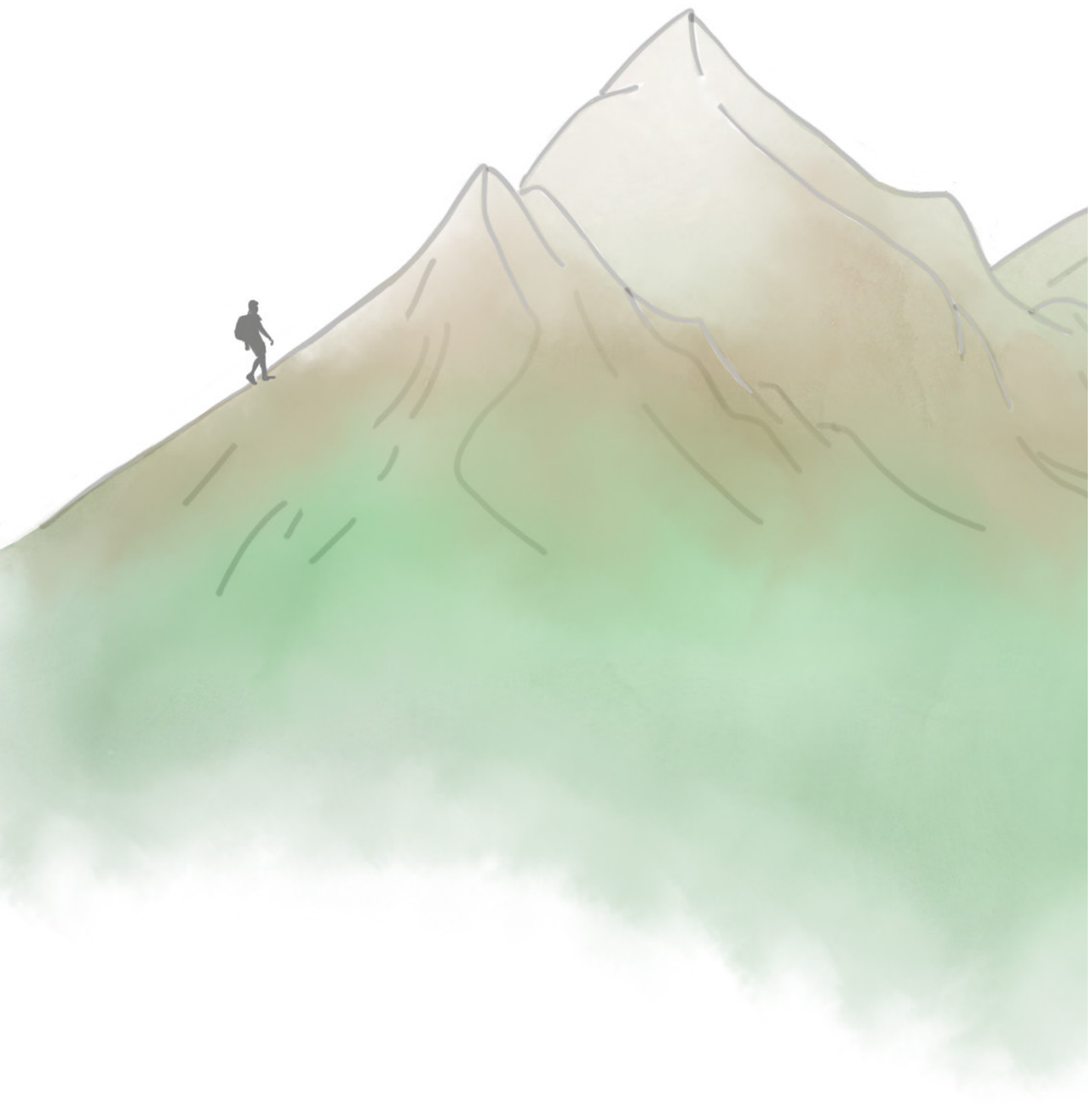
44 Austin, P. C., Harrell, F. E., Jr. & van Klaveren, D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med* **39**, 2714-2742, doi:10.1002/sim.8570 (2020).

45 Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6, doi:10.1136/bmj.i6 (2016).

46 Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* **26**, 565-574, doi:10.1177/0272989X06295361 (2006).

47 Vickers, A. J., Cronin, A. M., Elkin, E. B. & Gonen, M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak* **8**, 53, doi:10.1186/1472-6947-8-53 (2008).

48 Kerr, K. F., Brown, M. D., Zhu, K. & Janes, H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J Clin Oncol* **34**, 2534-2540, doi:10.1200/JCO.2015.65.5654 (2016).



Chapter 2

Prediction and clinical utility of a contralateral breast cancer risk model

Breast Cancer Research. 2019 Dec; 21(1):1-3[#]
https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-019-1221-1

Daniele Giardiello
Ewout W. Steyerberg,
Michael Hauptmann,
Muriel A. Adank, Delal Akdeniz, Carl Blomqvist, Stig E. Bojesen, Manjeet K. Bolla, Mariël Brinkhuis, Jenny Chang-Claude, Kamila Czene, Peter Devilee, Alison M. Dunning, Douglas F. Easton, Diana M. Eccles, Peter A. Fasching, Jonine Figueroa, Henrik Flyger, Montserrat García-Closas, Lothar Haeberle, Christopher A. Haiman, Per Hal, Ute Hamann, John L. Hopper, Agnes Jager, Anna Jakubowska, Audrey Jung, Renske Keeman, Iris Kramer, Diether Lambrechts, Loic Le Marchand, Annika Lindblom, Jan Lubiński, Mehdi Manoochehri, Luigi Mariani, Heli Nevanlinna, Hester S.A. Oldenburg, Saskia Pelders, Paul D.P. Pharoah, Mitul Shah, Sabine Siesling, Vincent T.H.B.M. Smit, Melissa C. Southey, William J. Tapper, Rob A.E.M. Tollenaar, Alexandra J. van den Broek, Carolien H.M. van Deurzen, Flora E. van Leeuwen, Chantal van Ongeval, Laura J. Van't Veer, Qin Wang, Camilla Wendt, Pieter J. Westenend,
Maartje J. Hooning
Marjanka K. Schmidt

[#]Author affiliations available on the journal's website

ABSTRACT

Background

Breast cancer survivors are at risk for contralateral breast cancer (CBC), with the consequent burden of further treatment and potentially less favorable prognosis. We aimed to develop and validate a CBC risk prediction model, and evaluate its applicability for clinical decision-making.

Methods

We included data of 132,756 invasive non-metastatic breast cancer patients from 20 studies with 4,682 CBC events and a median follow-up of 8.8 years. We developed a multivariable Fine and Gray prediction model (PredictCBC-1A) including patient, primary tumor, and treatment characteristics, and *BRCA1/2* germline mutation status, accounting for the competing risks of death and distant metastasis. We also developed a model without *BRCA1/2* mutation status (PredictCBC-1B) since this information was available for only 6% of patients and is routinely unavailable in the general breast cancer population. Prediction performance was evaluated using calibration and discrimination, calculated by a time-dependent Area-Under-the-Curve (AUC) at 5 and 10 years after diagnosis of primary breast cancer, and an internal-external cross-validation procedure. Decision curve analysis was performed to evaluate the net benefit of the model to quantify clinical utility.

Results

In the multivariable model, *BRCA1/2* germline mutation status, family history and systemic adjuvant treatment showed the strongest associations with CBC risk. The AUC of PredictCBC-1A was 0.63 (95% prediction interval (PI) at 5 years: 0.52–0.74; at 10 years: 0.53–0.72). Calibration in-the-large was -0.13 (95%PI: -1.62–1.37) and the calibration slope was 0.90 (95%PI: 0.73–1.08). The AUC of Predict-1B at 10 years was 0.59 (95% PI: 0.52–0.66); calibration was slightly lower. Decision curve analysis for preventive contralateral mastectomy showed potential clinical utility of PredictCBC-1A between thresholds of 4-10% 10-year CBC risk for *BRCA1/2* mutation carriers and non-carriers.

Conclusions

We developed a reasonably calibrated model to predict the risk of CBC in women of European-descent, however, prediction accuracy was moderate. Our model shows potential for improved risk counseling, but decision making regarding contralateral preventive mastectomy, especially in the general breast cancer population where limited information of the mutation status in *BRCA1/2* is available, remains challenging.

INTRODUCTION

Breast cancer (BC) is a major burden for women's health[1]. Survival has improved substantially over the past half century due to earlier detection and advanced treatment modalities, for example in the Netherlands, 10-year survival of a first primary BC improved from 40% in 1961-1970 to 79% in 2006-2010[2]. Consequently, an increasing numbers of BC survivors are at risk to develop a new primary tumor in the opposite (contralateral) breast, with subsequent treatment and potentially less favorable prognosis[3]. BC survivors are more likely to develop contralateral breast cancer (CBC) compared to healthy women to develop a first primary BC[4].

Women at elevated CBC risk have been identified to be *BRCA1/2* and *CHEK2* c.1100del mutation carriers and to have a BC family history, particular a family history of bilateral BC[5-10]. For *BRCA1/2* mutation carriers, in whom CBC risk is high, contralateral preventive mastectomy (CPM) is often offered[11]. However, the average risk of CBC among all first BC survivors is still relatively low, with an incidence of ~0.4% per year[12-14]. Despite this, in recent years, CPM frequency has increased among women in whom CBC risk is low[15]. For these reasons, there is an urgent need for improved individualized prediction of CBC risk, both to facilitate shared-decision making of physicians and women regarding treatment and prevention strategies for those at high CBC risk and to avoid unnecessary CPM or surveillance mammography after first primary BC when CBC risk is low.

To our knowledge, only one specific CBC risk prediction model (CBCrisk) has been developed to date. CBCrisk used data on 1,921 CBC cases and 5,763 matched controls with validation in two independent US studies containing a mix of invasive and *in-situ* BC[16, 17]. Moreover, the level of prediction performance measures such as calibration and discrimination needed for a CBC risk prediction to be clinically useful have not yet been addressed[18].

Our aim was two-fold: first, to develop and validate a CBC risk prediction model using a large international series of individual patient data including 132,756 patients with a first primary invasive BC between 1990 and 2013 from multiple studies in Europe, US and Australia with 4,682 incident CBCs; and second, to evaluate the potential clinical utility of the model to support decision making.

MATERIAL AND METHODS

Study population

We used data from five main sources: three studies from the Netherlands, 16 studies from

the Breast Cancer Association Consortium (BCAC), and a cohort from the Netherlands Cancer Registry[19-22]. For details regarding data collection and patient inclusion see **Supplementary Material section: Data and patient selection** and **TableS1**, and **Table S2**. We included female patients with invasive non-metastatic first primary BC with no prior history of cancer (except for non-melanoma skin cancer). The studies were either population- or hospital-based series; most women were of European-descent. We only included women diagnosed after 1990 to have a population with diagnostic and treatment procedures likely close to modern practice and at the same time sufficient follow-up to study CBC incidence; in total 132,756 women from 20 studies were included. All studies were approved by the appropriate ethics and scientific review boards. All women provided written informed consent or did not object to secondary use of clinical data in accordance with Dutch legislation and codes of conduct[23, 24].

Available data and variable selection

Several factors have been shown or suggested to be associated with CBC risk, including age at first BC, family history for BC, *BRCA1/2* and *CHEK2* c.1100del mutations, body mass index (BMI), breast density change, (neo)adjuvant chemotherapy, endocrine therapy, CPM, and characteristics of the first BC such as histology (lobular vs ductal), estrogen receptor (ER) status, lymph node status, tumor size, and TNM stage [5, 9, 12, 25-36]. The choice of factors to include in the analyses was determined by evidence from literature, availability of data in the cohorts, and current availability in clinical practice. We extracted the following information: *BRCA1/2* germline mutation, (first degree) family history of primary BC, and regarding primary BC diagnosis: age, nodal status, size, grade, morphology, ER status, progesterone-receptor (PR), human epidermal growth factor receptor 2 (HER2) status, administration of adjuvant and/or neoadjuvant chemotherapy, adjuvant endocrine therapy, adjuvant trastuzumab therapy, radiotherapy. We excluded PR status and TNM stage of the primary BC due to collinearity with ER status and the size of the primary tumor, respectively. In the current clinical practice, only patients with ER-positive tumors receive endocrine therapy and only patients with HER2-positive tumors receive trastuzumab; these co-occurrences were considered in the model by using composite categorical variables. More information is available online about the factors included in the analyses **(Supplementary Material: Data patient selection** and **Figure S1)**, follow-up per dataset, and study design **(Table S2)**.

Statistical analyses

All analyses were performed using SAS (SAS Institute Inc., Cary, NC, USA) and R software[37].

Primary endpoint, follow-up and predictors

The primary endpoint in the analyses was *in-situ* or invasive metachronous CBC. Follow-up started three months after invasive first primary BC diagnosis, in order to exclude synchronous CBCs, and ended at date of CBC, distant metastasis (but not at loco-regional relapse), CPM, or last date of follow-up (due to death, being lost to follow-up, or end of study), whichever occurred first. The follow-up of 27,155 (20.4%) women from the BCAC studies, recruited more than 3 months after diagnosis of the first primary BC (prevalent cases) started at recruitment (left truncation). Distant metastasis and death due to any cause were considered as competing events. Patients who underwent CPM during the follow-up were censored because the CBC risk was almost zero after a CPM[38].Missing data were multiply imputed by chained equations (MICE) to avoid loss of information due to case-wise deletion[39, 40]. Details about the imputation model, strategy used, and the complete case analysis, are provided in the **Supplementary Material: Multiple Imputation of missing values, Complete case analysis,** and **Model diagnostics and baseline recalibration** and **Tables S3 and S4**

Model development and validation

For model development, we used a multivariable Fine and Gray model regression to account for death and distant metastases as competing events[41, 42]. Heterogeneity of baseline risks between studies was taken into account using the study as a stratification term. A stratified model allows the baseline subdistribution hazard to be different across the studies and parameter estimation is performed by maximization of the partial likelihood per study. A Breslow-type estimator was used to estimate the baseline cumulative subdistribution hazard per study. The assumption of proportional subdistribution hazards was graphically checked using Schoenfield residuals[43]. The resulting subdistributional hazard ratios (sHRs) and corresponding 95% confidence intervals (CI) were pooled from the 10 imputed data sets using Rubin's rules[44]. We built a nomogram for estimating the 5- and 10- year cumulative incidence of CBC as a graphical representation of the multivariable risk prediction model[45].

The validity of the model was investigated by leave-one-study-out cross-validation; i.e., in each validation step all studies are used except one in which the validity of the model is evaluated[46, 47]. Since the ABCS study and some studies from BCAC had insufficient CBC events required for reliable validation, we used the geographic area as unit of splitting. We had 20 studies in five main sources: 17 out of 20 studies that were combined in 4 geographic areas. In total, 3 studies and 4 geographic areas were used to assess the prediction performance of the model (see **Supplementary Material: Leave-one-study-out cross-validation** and **Table S5**)[47, 48].

The performance of the model was assessed by discrimination ability to differentiate between patients who experienced CBC and those who did not, and by calibration, which measures the agreement between observed and predicted CBC risk. Discrimination

was quantified by time-dependent Area under the ROC Curves (AUCs) based on Inverse Censoring Probability Weighting at 5 and 10 years [49, 50]. In presence of competing risks, the R package timeROC provides two types of AUC according to different definition of time-dependent cases and controls. AUCs were calculated considering a patient who developed a CBC as a case and a patient free of any event as a control at 5 and 10 years[50]. Values of AUCs close to 1 indicate good discriminative ability, while values close to 0.5 indicated poor discriminative ability. Calibration was assessed by the calibration-in-the-large and slope statistic[51]. Calibration-in-the-large lower or higher than zero indicates that prediction is systematically too high or low, respectively. A calibration slope of 1.0 indicates good overall calibration; slopes below (above) 1.0 indicate over (under) estimation of risk by the model.

To allow for heterogeneity among studies, a random-effect meta-analysis was performed to provide summaries of discrimination and calibration performance. The 95% prediction intervals (PI) indicated the likely range for the prediction performances of the model in a new dataset. Further details about the validation process are provided in **Supplementary Leave-one-study-out cross-validation**.

Clinical utility

The clinical utility of the prediction model was evaluated using decision curve analysis (DCA)[52, 53]. Such a decision may apply to more or less intensive screening and follow-up or to decision of a CPM. The key part of the DCA is the net benefit, which is the number of true-positive classifications (in this example: the benefit of CPM to a patient who would have developed a CBC) minus the number of false-positive classifications (in this example: the harm of unnecessary CPM in a patient who would not have developed a CBC). The false-positives are weighted by a factor related to the relative harm of a missed CBC versus an unnecessary CPM. The weighting is derived from the threshold probability to develop a CBC using a defined landmark time point (e.g. CBC risk at 5 or 10 years)[54]. For example, a threshold of 10% implies that CPM in 10 patients, of whom one would develop CBC if untreated, is acceptable (thus performing 9 unnecessary CPMs). The net benefit of a prediction model is traditionally compared with the strategies of treat all or treat none. Since the use of CPM is generally only suggested among *BRCA1/2* mutation carriers, for a more realistic illustration the decision curve analysis was reported among *BRCA1/2* mutation carriers and non-carriers[55]. See **Supplementary material: Clinical utility** for details.

RESULTS

A total of 132,756 invasive primary BC women diagnosed between 1990 and 2013, with 4,682 CBC events, from 20 studies, were used to derive the model for CBC risk (**Table S2**). Median follow-up time was 8.8 years and CBC cumulative incidences at 5 and 10 years were 2.1% and 4.1%, respectively. Details of the studies and patient, tumor, and treatment characteristics are provided in **Table S6**. The multivariable model with estimates for all included factors is shown in **Table 1**. *BRCA1/2* germline mutation status, family history and systemic adjuvant treatment showed the strongest associations with CBC risk.

Table 1. Multivariable subdistribution hazard model for contralateral breast cancer risk

Factor (category) at primary breast cancer	Multivariable analysis	
	sHR	95% CI
Age, *years*	0.68*	0.62 - 0.74*
Family history (yes versus no)	1.35	1.27 - 1.45
BRCA mutation		
BRCA1 versus non-carrier	3.68	3.34 - 4.07
BRCA2 versus non-carrier	2.56	2.36 - 2.78
Nodal status (positive versus negative)	0.87	0.80 - 0.93
Tumor size, *cm*		
(2,5] versus ≤ 2	0.95	0.89 - 1.02
> 5 versus ≤ 2	1.14	0.99 - 1.31
Morphology (lobular including mixed versus ductal including other)	1.23	1.14 - 1.34
Grade		
Moderately differentiated versus well differentiated	0.89	0.82 - 0.96
Poorly differentiated versus well differentiated	0.75	0.70 - 0.82
Chemotherapy (yes versus no)	0.77	0.70 - 0.84
Radiotherapy to the breast (yes versus no)	1.01	0.95 - 1.08
ER (positive or negative) / endocrine therapy (yes or no)		
Negative/no versus positive/yes	1.43	1.30 - 1.57
Positive/no versus positive/yes	1.75	1.61 - 1.90
HER2 (positive or negative) / trastuzumab therapy (yes or no)		
Negative/no versus positive/yes	1.08	0.93 - 1.27
Positive/no versus positive/yes	0.99	0.83 - 1.18

Abbreviations: sHR: subdistributional hazard ratio; CI: confidence interval; ER: estrogen receptor; HER2: human epidermal growth factor receptor 2; *Age was parameterized as a linear spline with one interior knot at 50 years. For representation purposes, we here provide the sHR for the 75th versus the 25th percentile. For more details about age parameterization, see also Supplementary Methods.

The prediction performance of the main model (PredictCBC, version 1A) based on the leave-one-study-out cross-validation method is shown in **Figure 1**. The AUC at 5 years was 0.63 (95% confidence interval (CI): 0.58–0.67; 95% prediction interval (PI): 0.52–0.74)); the AUC at 10 years was also 0.63 (95%CI: 0.59–0.66; 95%PI: 0.53–0.72). Calibrations showed some indications of overestimation of risk. The calibration-in-

the-large was -0.13 (95%CI: -0.66–0.40; 95%PI: -1.62–1.37). The calibration slope was 0.90 (95%CI: 0.79–1.02; 95%PI: 0.73–1.08) in the cross-validation. Calibration plots are provided in **Figure S2 and S3**.

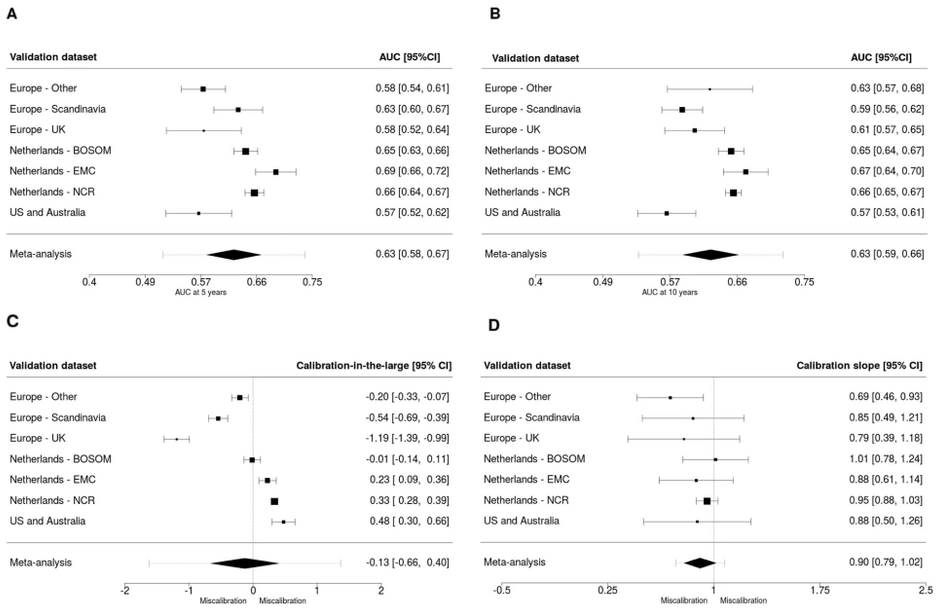


Figure 1. Analysis of predictive performance in leave-one-study-out cross-validation.
Panel A and B show the discrimination assessed by a time-dependent AUC at 5 and 10 years, respectively. Panel C shows the calibration accuracy measured with calibration in-the-large. Panel D shows the calibration accuracy measured with calibration slope. The black squares indicate the estimated accuracy of a model built using all remaining studies or geographic areas. The black horizontal lines indicate the corresponding 95% confidence intervals of the estimated accuracy (interval whiskers). The black diamonds indicate the mean with the corresponding 95% confidence intervals of the predictive accuracy and the dashed horizontal lines indicate the corresponding 95% prediction intervals.

The nomogram representing a graphical tool for estimating the CBC cumulative incidence at 5 and 10 years based on our model and the estimated baseline of the Dutch Cancer Registry is shown in **Figure 2**. In the nomogram, the categories of each factor are assigned a score using the topmost 'Points' scale, then all scores are summed up to obtain the 'Total points', which relate to the cumulative incidence of CBC. The formulae of the models (PredictCBC-1A and 1B) providing the predicted cumulative incidence are given in **Supplementary Material: Formula to estimate the CBC risk** and **Formula to estimate CBC risk in patients not tested for *BRCA***.

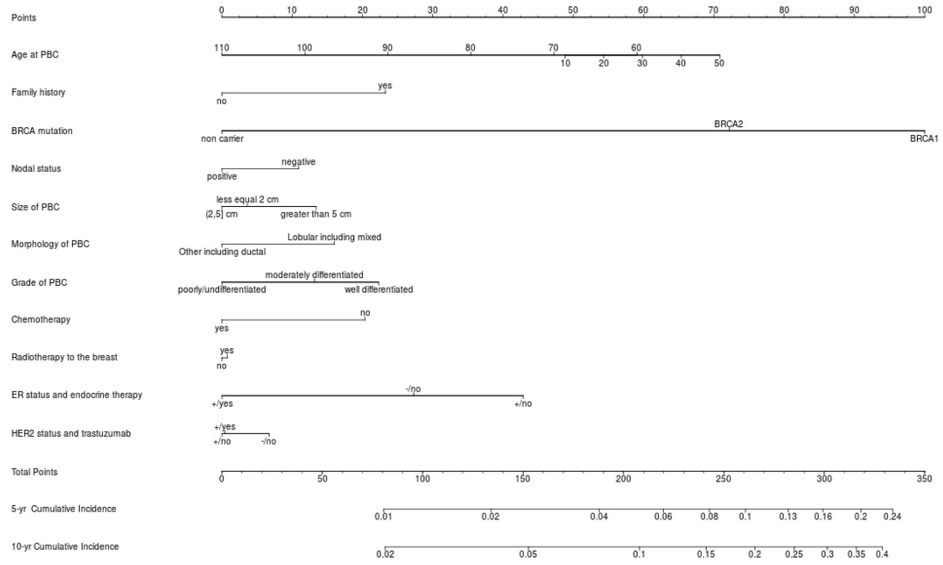


Figure 2. Nomogram for prediction of 5- and 10-year contralateral breast cancer cumulative incidence.
The 5- and 10-years contralateral breast cancer cumulative incidence is calculated by taking the sum of the risk points, according to patient, first primary breast cancer tumor, and treatment characteristics. For each factor, the number of associated risk points can be determined by drawing a vertical line straight up from the factor's corresponding value to the axis with risk points (0-100). The total points axis (0-350) is the sum of the factor's corresponding values determined by every individual patient's characteristics. Draw a line straight down from the total points axis to find the 5- and 10-years cumulative incidence.
PBC=primary breast cancer; ER=estrogen receptor status; HER2= human epidermal growth factor receptor 2; yr=year

The DCAs for preventive contralateral mastectomy showed potential clinical utility of PredictCBC-1A between thresholds of 4-10% 10-year CBC risk for *BRCA1/2* mutation carriers and non-carriers (**Table 2**). For example, if we find it acceptable that one in 10 patients for whom a CPM is recommended develops a CBC, a risk threshold of 10% may be used to define high and low risk *BRCA1/2* mutation carriers based on the absolute 10-year CBC risk prediction estimated by the model. Compared with a strategy recommending CPM to all carriers of a mutation in *BRCA1*/2, this strategy avoids 161 CPMs per 1,000 patients. In contrast, almost no non *BRCA1/2* mutation carriers reach the 10% threshold (the general BC population, **Figure 3**). The decision curves provide a comprehensive overview of the net benefit for a range of harm-benefit thresholds at 10-year CBC risk (**Figure 4**).

Decision curves for CBC risk at 5 year and the corresponding clinical utility are provided in **Figure S4** and **Table S7**, respectively.

Table 2: Clinical utility of the 10-year contralateral breast cancer risk prediction model. At the same probability threshold, the net benefit is exemplified in *BRCA1/2* mutation carriers (for avoiding unnecessary CPM) and non-carriers (performing necessary CPM).

		BRCA1/2 mutation carriers		Non-carriers	
Probability threshold p_t (%)	Unnecessary CPMs needed to prevent a CBC*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients
4	24.0	0.0	0.0	3.9	93.6
5	19.0	0.0	0.0	2.1	39.9
6	15.7	0.1	1.6	0.5	7.8
7	13.3	1.9	25.2	0.1	1.3
8	11.5	5.5	63.3	0.0	0.0
9	10.1	10.7	108.2	0.0	0.0
10	9.0	17.9	161.1	0.0	0.0

CPM: contralateral preventive mastectomy; CBC: contralateral breast cancer;
*The number of unnecessary contralateral mastectomies needed to prevent a CBC is calculated by: $(1-p_t)/p_t$. See also Supplementary Methods.

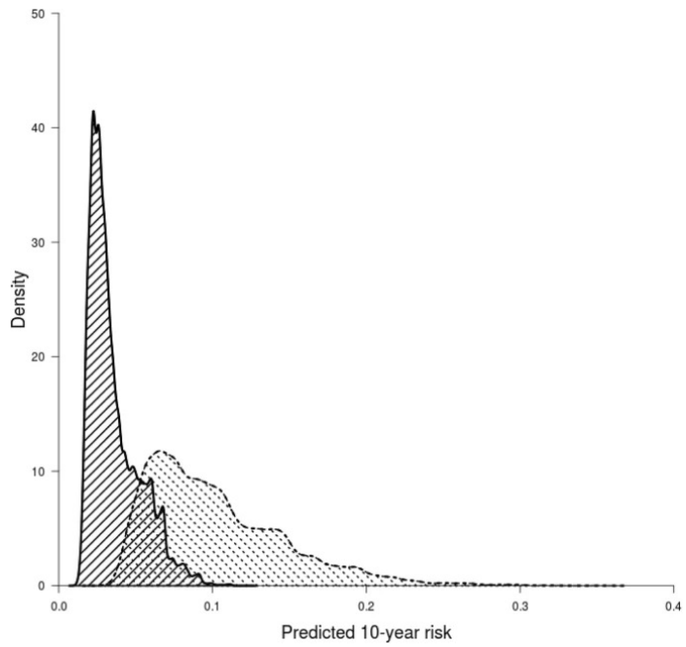
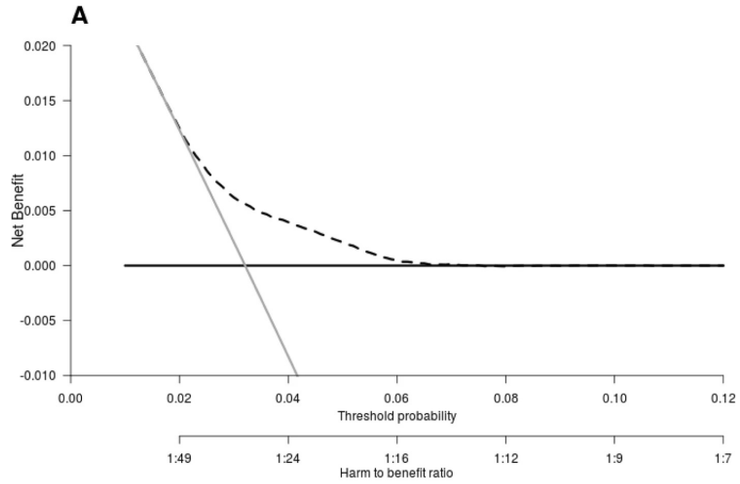


Figure 3: Density distribution of 10-year predicted contralateral breast cancer absolute risk within non-carriers (area with black solid lines) and *BRCA1/2* mutation carriers (area with black dashed lines).



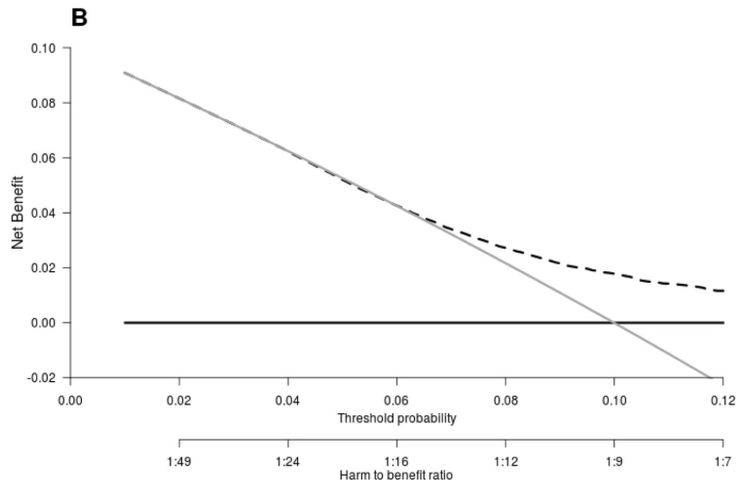


Figure 4: Decision curve analysis at 10 years for the contralateral breast cancer risk model including *BRCA* mutation information.
Panel A shows the decision curve to determine the net benefit of the estimated 10-year predicted contralateral breast cancer (CBC) cumulative incidence for patients without a *BRCA1/2* gene mutation using the prediction model (dotted black line) compared to not treating any patients with contralateral preventive mastectomy (CPM) (black solid line). Panel B shows the decision curve to determine the net benefit of the estimated 10-year predicted CBC cumulative incidence for *BRCA1/2* mutation carriers using the prediction model (dotted black line) versus treating (or at least counseling) all patients (grey solid line). The y-axis measures net benefit, which is calculated by summing the benefits (true positives, i.e., patients with a CBC who needed a CPM) and subtracting the harms (false positives, i.e., patients with CPM who do not need it). The latter are weighted by a factor related to the relative harm of a non-prevented CBC versus an unnecessary CPM. The factor is derived from the threshold probability to develop a CBC at 10 years at which a patient would opt for CPM (e.g. 10%). The x-axis represents the threshold probability. Using a threshold probability of 10% implicitly means that CPM in 10 patients of whom one would develop a CBC if untreated is acceptable (9 unnecessary CPMs, harm to benefit ratio 1:9).

We also derived a risk prediction model (PredictCBC, version 1B) omitting *BRCA* status to provide CBC risk estimates for first BC patients not tested for *BRCA1/2* mutations. This model has slightly lower prediction performance; AUC at 5 and 10 years was both 0.59 (at 5 years: 95% CI: 0.54–0.63; 95% PI: 0.46–0.71; at 10 years:0.56–0.62; 95% PI: 0.52–0.66), calibration-in-the-large was -0.17 (95% CI: -0.72–0.38; 95% PI: -1.70–1.36) and calibration slope was 0.81 (95% CI: 0.63–0.99; 95% PI: 0.50–1.12) (**Supplementary Material Results of the prediction model without *BRCA* mutation**). Details of development, validation, and clinical utility are provided in **Tables S8-10** and **Figures S5-10**.

In a sensitivity analysis (see **Supplementary Material: Assessment of limited information of CPM**), we studied the impact of CPM on our results using two studies, in which CPM information was (almost) completely available. The lack of CPM information on cumulative incidence estimation hardly affected the results of our analyses (**Figure S11**).

DISCUSSION

Using established risk factors for CBC which are currently available in clinical practice, we developed PredictCBC, which can be used to calculate 5- and 10-year absolute CBC risk. The risk prediction model includes carriership of *BRCA1/2* mutations, an important determinant of CBC risk in the decision-making process[6].

The calibration of the model was reasonable and discrimination moderate within the range of other tools commonly used for routing counseling and decision-making in clinical oncology for primary BC risk[56-59]. As expected, the prediction accuracy was lower when we omitted the *BRCA* mutation carrier status although the prevalence of *BRCA* mutations among BC patients is quite low (2-4%)[60, 61].

In the breast cancer population, CBC is a relatively uncommon event (~0.4% per year) and difficult to predict. Therefore physicians should carefully consider which patients should consider CPM using a prediction model[62]. The current clinical recommendations of CPM are essentially based on the presence of a mutation in the *BRCA1/2* genes. Based on the risk distribution defined by the current model (**Figure 3**), this is a reasonable approach: essentially no non-carrier women reach a 10% risk 10-year threshold. However, more than 50% of carriers do not reach this threshold either, suggesting that a significant proportion of *BRCA1/2* carriers might be spared CPM. Contralateral surveillance mammography may also be avoided although detection and knowledge of recurrences may be necessary for better defined individualized follow-up and patient-tailored treatment strategies[63, 64].

CBC risk patterns and factors were identified previously in a large population-based study with 10,944 CBC of 212,630 patients from the Surveillance, Epidemiology and End Results (SEER) database diagnosed from 1990 to 2013[65]. However, SEER does not include details of endocrine treatment and chemotherapy, therapies administrated to reduce recurrences and CBCs[13, 66]. Furthermore, in this study the model was not validated or evaluated based on prediction accuracy, nor was a tool provided. Another study provided general guidelines for CPM by calculating the life-time risk of CBC based on a published systematic review of age at first BC, *BRCA1/2* gene mutation, family history of BC, ER status, ductal carcinoma in situ, and oophorectomy[34, 67]. However, the authors specified that the calculation of the CBC life-time risk should be considered only as a guide for helping clinicians to stratify patients into risk categories rather than a precise tool for the objective assessment of the risk.

Only one other prediction model (CBCrisk) has been developed and validated using data of 1,921 CBC cases and 5,763 matched controls[16]. External validation of CBCrisk of two independent datasets using 5,185 and 6,035 patients with 111 and 117 CBC assessed a discrimination between 0.61 and 0.65[17]. The discrimination of our PredictCBC model at 5 and 10 years was similar, however the geographic diversity of the studies gave a more complete overview of external validity[47]. Moreover, we showed the net benefit of our model using decision curve analysis since standard performance metrics of discrimination, calibration, sensitivity, and specificity alone are insufficient to assess the clinical utility[18, 53].

Some limitations of our study must be recognized. First, reporting of CBC was not entirely complete in all studies and information about CPM was limited in most datasets, which may have underestimated the cumulative incidence, although the overall 10-year cumulative incidence of 4.1% is in line with other data[5, 34]. Second, some women included in the Dutch studies (providing specific information on family history, *BRCA* mutation or CPM) were also present in our selection of the Netherlands Cancer Registry population. Privacy and coding issues prevented linkage at the individual patient level, but based on the hospitals from which the studies recruited, and the age and period criteria used, we calculated a maximum potential overlap of 3.4%. Third, in the United States and Australian datasets, the prediction performance was uncertain due to limited sample size and missing values. Moreover, some important predictors such as family history and especially *BRCA* mutation status were only available in a subset of the women (from familial- and unselected hospital-based studies) and patients with data on *BRCA* mutation status might have been insufficiently represented for tested populations and further development and validation of PredictCBC-1A will be necessary. However, although *BRCA1/2* mutation information was unavailable in 94% of our data, the approach of the imputation led to consistently good performing models[68-70]. The remaining factors were

quite complete: ~79% of patients had at most one missing factor, which provided good imputation diagnostic performances. Since most BC patients are not currently tested in the clinical practice for *BRCA1/2* mutations, we assessed the clinical utility of PredictCBC version 1B to provide individualized CBC risk estimates for first BC patients not tested for *BRCA1/2* germline mutations [60, 71]. Our PredictCBC version 1B model provides less precise estimates, but may be useful in providing general CBC risk estimates, which could steer women away from CPM or trigger *BRCA* testing.

Last but not least, adequate presentation of the risk estimates from the PredictCBC-1A and PredictCBC-1B is crucial for effective communication about CBC risk during doctor-patient consultations [72, 73]. A nomogram is an important component to communicate the risk of modern medical decision making, although it may be difficult to use and might potentially make it more difficult to interpret the risks for laymen[74] An online tool is being implemented, and a pilot-study will be conducted amongst patients and clinicians to assess how the risk estimates from PredictCBC-1A and 1B can best be visualized to facilitate communication with patients. Other factors, which were not available in our study, predict breast cancer risk and their inclusion may further improve the discrimination and clinical utility of our CBC risk model: these factors include *CHEK2* c.1100del mutation carriers, polygenic risk scores based on common genetic variants, breast density, reproductive and life-style factors such as BMI and age at menarche[75]. Additional data with complete information of *BRCA1/2* mutation should be also considered in the model upgrade to reduce uncertainty of CBC risk estimates. External validation in other studies including patients of other ethnicities, will also be important. In the meantime, our model provides a reliable basis for CBC risk counseling.

CONCLUSIONS

In conclusion we have developed and cross-validated risk prediction models for CBC (PredictCBC) based on different European-descent population and hospital-based studies. The model is reasonably calibrated and prediction accuracy is moderate. The clinical utility assessment of PredictCBC showed potential for improved risk counseling, although decision regarding CPM in the general breast cancer population remains challenging. Similar results have been found for PredictCBC version 1B, a CBC risk prediction model that calculates individualized CBC risk for first BC patients not tested for *BRCA1/2* germline mutation.

Abbreviations

AUC: Area-under-the-ROC-curve; **BC:** Breast cancer; **BCAC:** Breast Cancer Association Consortium;**BMI:** Body mass index; **CBC:** Contralateral breast cancer; **CI:** Confidence interval; **CPM:** Contralateral preventive mastectomy; **DCA:** Decision curve analysis; **ER:** Estrogen receptor; **HER2:** Human epidermal growth receptor 2; **MICE:** Multiple imputation by chained equations; **PI:** Prediction interval; **PR:** Progesterone receptor; **SEER:** Surveillance, Epidemiology and End Results; **TNM:**TNM Classification of Malignant Tumors.

Acknowledgements

We thank all individuals who took part in these studies and all researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out.

ABCFS thank Maggie Angelakos, Judi Maskiell, Gillian Dite. ABCS and BOSOM thanks all the collaborating hospitals and pathology departments and many individual that made this study possible; specifically we wish to acknowledge: Annegien Broeks, Sten Cornelissen, Frans Hogervorst, Laura van 't Veer, Floor van Leeuwen, Emiel Rutgers. EMC thanks J.C. Blom-Leenheer, P.J. Bos,C.M.G. Crepin and M. van Vliet for data management. CGPS thanks staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. HEBCS thanks Taru A. Muranen, Kristiina Aittomäki, Karl von Smitten, Irja Erkkilä. KARMA thanks the Swedish Medical Research Counsel. LMBC thanks Gilian Peuteman, Thomas Van Brussel, EvyVanderheyden and Kathleen Corthouts. MARIE thanks Petra Seibold, Dieter Flesch-Janys, Judith Heinz, Nadia Obi, Alina Vrieling, Sabine Behrens, Ursula Eilber, Muhabbet Celik, Til Olchers and Stefan Nickels. ORIGO thanks E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The authors thank the registration team of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry as well as IKNL staff for scientific advice. PBCS thanks Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner. The ethical approval for the POSH study is MREC /00/6/69, UKCRN ID: 1137. We thank the SEARCH team.

Funding

This work is supported by the Alpe d'HuZes/Dutch Cancer Society (KWF Kankerbestrijding) project 6253.

BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report.

The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. BOSOM was supported by the Dutch Cancer Society grant numbers DCS-NKI 2001-2423, DCS-NKI 2007-3839, and DCSNKI 2009-4363; the Cancer Genomics Initiative; and notary office Spier & Hazenberg for the coding procedure. The EMC was supported by grants from Alpe d'HuZes/Dutch Cancer Society NKI2013-6253 and from Pink Ribbon 2012.WO39.C143. The HEBCS was financially supported by the Helsinki University Hospital Research Fund, the Finnish Cancer Society, and the Sigrid Juselius Foundation.

Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantzows foundation. The KARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. LMBC is supported by the 'Stichting tegen Kanker'. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419, 110826, 110828], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MEC was support by NIH grants CA63464, CA54281, CA098758, CA132839 and CA164973. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. Genotyping for PLCO was supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The POSH study is funded by Cancer Research UK (grants C1275/A11699, C1275/C22524, C1275/A19187, C1275/A15956 and Breast Cancer Campaign 2010PR62, 2013PR044. PROCAS is funded from NIHR grant PGfAR 0707-10031. SEARCH is funded by Cancer Research UK [C490/A10124, C490/A16561] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. SKKDKFZS is supported by the DKFZ. The SZBCS (Szczecin Breast Cancer Study) was supported by Grant PBZ_KBN_122/P05/2004 and The National Centre for Research and Development (NCBR) within the framework of the international ERA-NET TRANSAN JTC 2012 application no. Cancer 12-054 (Contract No. ERA-NET-TRANSCAN / 07/2014).

Availability of data and materials

All data relevant to this report are included in this published article and its supplementary information files. The datasets analyzed during the current study are not publicly available due to protection of participant privacy and confidentiality, and ownership of the contributing institutions, but may be made available in anonymized form via the corresponding author on reasonable request and after approval of the involved institutions.

Authors' contributions

MKS and MJH conceived the study in collaboration with EWS and MH. DG performed the statistical analysis. DG, MKS, MJH, EWS and MH interpreted the results and drafted the manuscript. MAA, DA, CB, SEB, MKB, MB, JCC, KC, PD,AMD, DFE, DME, PAF, JF,HF,MGC,LK, CAH, PH, UH, JLH, AG, AJ1, AJ2, RK, IK, DL, LLM, AL, JL, MM, LM, HN, HSAO, SP, PDPP, MS, SS, VTHBMS, MCS, WJT, RAEMT, AJvB, CHMvD, FEvL, CvO, LJvV, QW, CW, PJW contributed to critical revision and editing of the final version of the manuscript for publication. All authors were involved in data generation or provision, and read and approved the final manuscript.

Ethics approval and consent to participate

Each study was approved by its institutional ethical review board.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin* 2018, **68**(6):394-424.
2. **Survival and prevalence of cancer** [https://www.cijfersoverkanker.nl]
3. Schaapveld M, Visser O, Louwman WJ, Willemse PH, de Vries EG, van der Graaf WT, Otter R, Coebergh JW, van Leeuwen FE: **The impact of adjuvant therapy on contralateral breast cancer risk and the prognostic significance of contralateral breast cancer: a population based study in the Netherlands**. *Breast Cancer Res Treat* 2008, **110**(1):189-197.
4. Brenner DJ: **Contralateral second breast cancers: prediction and prevention**. *J Natl Cancer Inst* 2010, **102**(7):444-445.
5. van den Broek AJ, van 't Veer LJ, Hooning MJ, Cornelissen S, Broeks A, Rutgers EJ, Smit VT, Cornelisse CJ, van Beek M, Janssen-Heijnen ML *et al*: **Impact of Age at Primary Breast Cancer on Contralateral Breast Cancer Risk in BRCA1/2 Mutation Carriers**. *J Clin Oncol* 2016, **34**(5):409-418.
6. Malone KE, Begg CB, Haile RW, Borg A, Concannon P, Tellhed L, Xue S, Teraoka S, Bernstein L, Capanu M *et al*: **Population-based study of the risk of second primary contralateral breast cancer associated with carrying a mutation in BRCA1 or BRCA2**. *J Clin Oncol* 2010, **28**(14):2404-2410.
7. Evans DG, Ingham SL, Baildam A, Ross GL, Lalloo F, Buchan I, Howell A: **Contralateral mastectomy improves survival in women with BRCA1/2-associated breast cancer**. *Breast Cancer Res Treat* 2013, **140**(1):135-142.
8. Graeser MK, Engel C, Rhiem K, Gadzicki D, Bick U, Kast K, Froster UG, Schlehe B, Bechtold A, Arnold N *et al*: **Contralateral breast cancer risk in BRCA1 and BRCA2 mutation carriers**. *J Clin Oncol* 2009, **27**(35):5887-5892.
9. Weischer M, Nordestgaard BG, Pharoah P, Bolla MK, Nevanlinna H, Van't Veer LJ, Garcia-Closas M, Hopper JL, Hall P, Andrulis IL *et al*: **CHEK2*1100delC heterozygosity in women with breast cancer associated with early death, breast cancer-specific death, and increased risk of a second breast cancer**. *J Clin Oncol* 2012, **30**(35):4308-4316.
10. Kuchenbaecker KB, Hopper JL, Barnes DR, Phillips KA, Mooij TM, Roos-Blom MJ, Jervis S, van Leeuwen FE, Milne RL, Andrieu N *et al*: **Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers**. *JAMA* 2017, **317**(23):2402-2416.
11. Domchek SM: **Risk-Reducing Mastectomy in BRCA1 and BRCA2 Mutation Carriers: A Complex Discussion**. *JAMA* 2019, **321**(1):27.
12. Chen Y, Thompson W, Semenciw R, Mao Y: **Epidemiology of contralateral breast cancer**. *Cancer Epidemiol Biomarkers Prev* 1999, **8**(10):855-861.
13. Kramer I, Schaapveld M, Oldenburg HSA, Sonke GS, McCool D, van Leeuwen FE, Van de Vijver KK, Russell NS, Linn SC, Siesling S *et al*: **The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype**. *J Natl Cancer Inst* 2019.
14. Portschy PR, Abbott AM, Burke EE, Nzara R, Marmor S, Kuntz KM, Tuttle TM: **Perceptions of Contralateral Breast Cancer Risk: A Prospective, Longitudinal Study**. *Ann Surg Oncol* 2015, **22**(12):3846-3852.
15. Murphy JA, Milner TD, O'Donoghue JM: **Contralateral risk-reducing mastectomy in sporadic breast cancer**. *Lancet Oncol* 2013, **14**(7):e262-269.
16. Chowdhury M, Euhus D, Onega T, Biswas S, Choudhary PK: **A model for individualized risk prediction of contralateral breast cancer**. *Breast Cancer Res Treat* 2017, **161**(1):153-160.
17. Chowdhury M, Euhus D, Arun B, Umbricht C, Biswas S, Choudhary P: **Validation of a personalized risk prediction model for contralateral breast cancer**. *Breast Cancer Res Treat* 2018.
18. Vickers AJ, Van Calster B, Steyerberg EW: **Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests**. *BMJ* 2016, **352**:i6.
19. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A *et al*: **Association analysis identifies 65 new breast cancer risk loci**. *Nature* 2017, **551**(7678):92-94.
20. Schmidt MK, Tollenaar RA, de Kemp SR, Broeks A, Cornelisse CJ, Smit VT, Peterse JL, van Leeuwen FE, Van't Veer LJ: **Breast cancer survival and tumor characteristics in premenopausal women carrying the CHEK2*1100delC germline mutation**. *J Clin Oncol* 2007, **25**(1):64-69.
21. Schmidt MK, van den Broek AJ, Tollenaar RA, Smit VT, Westenend PJ, Brinkhuis M, Oosterhuis WJ, Wesseling J, Janssen-Heijnen ML, Jobsen JJ *et al*: **Breast Cancer Survival of BRCA1/BRCA2 Mutation Carriers in a Hospital-Based Cohort of Young Women**. *J Natl Cancer Inst* 2017, **109**(8).
22. Font-Gonzalez A, Liu L, Voogd AC, Schmidt MK, Roukema JA, Coebergh JW, de Vries E, Soerjomataram I: **Inferior survival for young patients with contralateral compared to unilateral breast cancer: a nationwide population-based study in the Netherlands**. *Breast Cancer Res Treat* 2013, **139**(3):811-819.
23. Riegman PH, van Veen EB: **Biobanking residual tissues**. *Hum Genet* 2011, **130**(3):357-368.
24. Foundation Federation of Dutch Medical Scientific Societies: **Human Tissue and Medical Research: Code of Conduct for responsible use**. 2011.
25. Vichapat V, Garmo H, Holmqvist M, Liljegren G, Warnberg F, Lambe M, Fornander T, Adolfsson J, Luchtenborg M, Holmberg L: **Tumor stage affects risk and prognosis of contralateral breast cancer: results from a large Swedish-population-based study**. *J Clin Oncol* 2012, **30**(28):3478-3485.
26. Vichapat V, Gillett C, Fentiman IS, Tutt A, Holmberg L, Luchtenborg M: **Risk factors for metachronous contralateral breast cancer suggest two aetiological pathways**. *Eur J Cancer* 2011, **47**(13):1919-1927.
27. Mariani L, Coradini D, Biganzoli E, Boracchi P, Marubini E, Pilotti S, Salvadori B, Silvestrini R, Veronesi U, Zucali R *et al*: **Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension**. *Breast Cancer Res Treat* 1997, **44**(2):167-178.
28. Reiner AS, Lynch CF, Sisti JS, John EM, Brooks JD, Bernstein L, Knight JA, Hsu L, Concannon P, Mellemkjaer L *et al*: **Hormone receptor status of a first primary breast cancer predicts contralateral breast cancer risk in the WECARE study population**. *Breast Cancer Res* 2017, **19**(1):83.
29. Sisti JS, Bernstein JL, Lynch CF, Reiner AS, Mellemkjaer L, Brooks JD, Knight JA, Bernstein L, Malone KE, Woods M *et al*: **Reproductive factors, tumor estrogen receptor status and contralateral breast cancer risk: results from the WECARE study**. *Springerplus* 2015, **4**:825.
30. Healey EA, Cook EF, Orav EJ, Schnitt SJ, Connolly JL, Harris JR: **Contralateral breast cancer: clinical characteristics and impact on prognosis**. *J Clin Oncol* 1993, **11**(8):1545-1552.
31. Gao X, Fisher SG, Emami B: **Risk of second primary cancer in the contralateral breast in women treated**

for early-stage breast cancer: a population-based study. *Int J Radiat Oncol Biol Phys* 2003, **56**(4):1038-1045.

32. Brooks JD, John EM, Mellemkjaer L, Lynch CF, Knight JA, Malone KE, Reiner AS, Bernstein L, Liang X, Shore RE *et al*: **Body mass index, weight change, and risk of second primary breast cancer in the WECARE study: influence of estrogen receptor status of the first breast cancer**. *Cancer Med* 2016, **5**(11):3282-3291.
33. Knight JA, Blackmore KM, Fan J, Malone KE, John EM, Lynch CF, Vachon CM, Bernstein L, Brooks JD, Reiner AS *et al*: **The association of mammographic density with risk of contralateral breast cancer and change in density with treatment in the WECARE study**. *Breast Cancer Res* 2018, **20**(1):23.
34. Basu NN, Barr L, Ross GL, Evans DG: **Contralateral risk-reducing mastectomy: review of risk factors and risk-reducing strategies**. *Int J Surg Oncol* 2015, **2015**:901046.
35. Akdeniz D, Schmidt MK, Seynaeve CM, McCool D, Giardiello D, van den Broek AJ, Hauptmann M, Steyerberg EW, Hooning MJ: **Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis**. *Breast* 2018, **44**:1-14.
36. Edge SB, Compton CC: **The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM**. *Ann Surg Oncol* 2010, **17**(6):1471-1474.
37. R Developement Core Team: **R: A Language and Environment for Statistical Computing**. In.: R: Foundation for Statistical Computing; 2017.
38. van den Broek AJ, Schmidt MK, van 't Veer LJ, Oldenburg HSA, Rutgers EJ, Russell NS, Smit V, Voogd AC, Koppert LB, Siesling S *et al*: **Prognostic Impact of Breast-Conserving Therapy Versus Mastectomy of BRCA1/2 Mutation Carriers Compared With Noncarriers in a Consecutive Series of Young Breast Cancer Patients**. *Ann Surg* 2019, **270**(2):364-372.
39. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group P-IS: **Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data**. *Stat Med* 2013, **32**(28):4890-4905.
40. Buuren Sv: **Flexible imputation of missing data**. Boca Raton, FL: CRC Press; 2012.
41. Geskus RB: **Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring**. *Biometrics* 2011, **67**(1):39-49.
42. Fine JP, Gray RJ: **A Proportional Hazards Model for the Subdistribution of a Competing Risk**. *Journal of the American Statistical Association* 1999, **94**(446):496-509.
43. Schoenfeld DA: **Sample-size formula for the proportional-hazards regression model**. *Biometrics* 1983, **39**(2):499-503.
44. Little RJA, Rubin DB: **Statistical analysis with missing data**. New York, N.Y.: Wiley; 1987.
45. Zhang Z, Geskus RB, Kattan MW, Zhang H, Liu T: **Nomogram for survival analysis in the presence of competing risks**. *Ann Transl Med* 2017, **5**(20):403.
46. Steyerberg EW, Harrell FE, Jr.: **Prediction models need appropriate internal, internal-external, and external validation**. *J Clin Epidemiol* 2016, **69**:245-247.
47. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW: **Geographic and temporal validity of prediction models: different approaches were useful to examine model performance**. *J Clin Epidemiol* 2016, **79**:76-85.
48. Collins GS, Ogundimu EO, Altman DG: **Sample size considerations for the external validation of a multivariable prognostic model: a resampling study**. *Stat Med* 2016, **35**(2):214-226.
49. Steyerberg EW: **Clinical prediction models : a practical approach to development, validation and updating**. New York: Springer; 2010.
50. Blanche P, Dartigues JF, Jacqmin-Gadda H: **Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks**. *Stat Med* 2013, **32**(30):5381-5397.
51. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, Riley RD: **Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model**. *J Clin Epidemiol* 2016, **69**:40-50.
52. Vickers AJ, Elkin EB: **Decision curve analysis: a novel method for evaluating prediction models**. *Med Decis Making* 2006, **26**(6):565-574.
53. Kerr KF, Brown MD, Zhu K, Janes H: **Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use**. *J Clin Oncol* 2016, **34**(21):2534-2540.
54. Vickers AJ, Cronin AM, Elkin EB, Gonen M: **Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers**. *BMC Med Inform Decis Mak* 2008, **8**:53.
55. Heemskerk-Gerritsen BA, Rookus MA, Aalfs CM, Ausems MG, Collee JM, Jansen L, Kets CM, Keymeulen KB, Koppert LB, Meijers-Heijboer HE *et al*: **Improved overall survival after contralateral risk-reducing mastectomy in BRCA1/2 mutation carriers with a history of unilateral breast cancer: a prospective analysis**. *Int J Cancer* 2015, **136**(3):668-677.
56. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA: **Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention**. *J Natl Cancer Inst* 2001, **93**(5):358-366.
57. Elmore JG, Fletcher SW: **The risk of cancer risk prediction: "What is my risk of getting breast cancer"?** *J Natl Cancer Inst* 2006, **98**(23):1673-1675.
58. Wishart GC, Azzato EM, Greenberg DC, Rashbass J, Kearins O, Lawrence G, Caldas C, Pharoah PD: **PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer**. *Breast Cancer Res* 2010, **12**(1):R1.
59. Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, Shak S, Baehner FL, Ravdin PM, Davidson NE *et al*: **Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features**. *J Clin Oncol* 2008, **26**(25):4063-4071.
60. van den Broek AJ, de Ruiter K, van 't Veer LJ, Tollenaar RA, van Leeuwen FE, Verhoef S, Schmidt MK: **Evaluation of the Dutch BRCA1/2 clinical genetic center referral criteria in an unselected early breast cancer population**. *Eur J Hum Genet* 2015, **23**(5):588-595.
61. Gail MH, Pfeiffer RM: **Breast Cancer Risk Model Requirements for Counseling, Prevention, and Screening**. *J Natl Cancer Inst* 2018.
62. O'Donnell M: **Estimating Contralateral Breast Cancer Risk**. *Current Breast Cancer Reports* 2018, **10**(2):91-97.
63. van Maaren MC, de Munck L, Strobbe LJA, Sonke GS, Westenend PJ, Smidt ML, Poortmans PMP, Siesling S: **Ten-year recurrence rates for breast cancer subtypes in the Netherlands: A large population-based**

study. *Int J Cancer* 2019, **144**(2):263-272.

64. Lu W, Schaapveld M, Jansen L, Bagherzadegan E, Sahinovic MM, Baas PC, Hanssen LM, van der Mijle HC, Brandenburg JD, Wiggers T *et al*: **The value of surveillance mammography of the contralateral breast in patients with a history of breast cancer**. *Eur J Cancer* 2009, **45**(17):3000-3007.
65. Xiong Z, Yang L, Deng G, Huang X, Li X, Xie X, Wang J, Shuang Z, Wang X: **Patterns of Occurrence and Outcomes of Contralateral Breast Cancer: Analysis of SEER Data**. *J Clin Med* 2018, **7**(6).
66. Langballe R, Mellemkjaer L, Malone KE, Lynch CF, John EM, Knight JA, Bernstein L, Brooks J, Andersson M, Reiner AS *et al*: **Systemic therapy for breast cancer and risk of subsequent contralateral breast cancer in the WECARE Study**. *Breast Cancer Res* 2016, **18**(1):65.
67. Basu NN, Ross GL, Evans DG, Barr L: **The Manchester guidelines for contralateral risk-reducing mastectomy**. *World J Surg Oncol* 2015, **13**:237.
68. Nieboer D, Vergouwe Y, Ankerst DP, Roobol MJ, Steyerberg EW: **Improving prediction models with new markers: a comparison of updating strategies**. *BMC Med Res Methodol* 2016, **16**(1):128.
69. Collins GS, Altman DG: **Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2**. *BMJ* 2012, **344**:e4181.
70. Madley-Dowd P, Hughes R, Tilling K, Heron J: **The proportion of missing data should not be used to guide decisions on multiple imputation**. *J Clin Epidemiol* 2019, **110**:63-73.
71. Childers CP, Childers KK, Maggard-Gibbons M, Macinko J: **National Estimates of Genetic Testing in Women With a History of Breast or Ovarian Cancer**. *J Clin Oncol* 2017, **35**(34):3800-3806.
72. Bonnett LJ, Snell KIE, Collins GS, Riley RD: **Guide to presenting clinical prediction models for use in clinical settings**. *BMJ* 2019, **365**:l737.
73. Van Belle V, Van Calster B: **Visualizing Risk Prediction Models**. *PLoS One* 2015, **10**(7):e0132614.
74. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP: **Nomograms in oncology: more than meets the eye**. *Lancet Oncol* 2015, **16**(4):e173-180.
75. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK *et al*: **Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes**. *Am J Hum Genet* 2019, **104**(1):21-34.

SUPPLEMENTARY MATERIALS

1. Data and patient selection

For this study we used data from five main sources available from national and international collaborations including nationwide registry data, as well as studies with more detailed information on relevant prediction factors[1-5]. Briefly, the five main sources were: (1) The Breast Cancer Association Consortium (BCAC), which is an international consortium of 102 studies comprising 182,898 patients (data version: January 2017) with a primary breast cancer (BC) diagnosed between 1939 and 2016[1]; (2) The Amsterdam Breast Cancer Study (ABCS) containing 2,390 patients diagnosed with a first BC at the Netherlands Cancer Institute – Antoni van Leeuwenhoek (NKI-AVL) hospital in Amsterdam from 2003 to 2013[2]; (3) The Breast Cancer Outcome Study of Mutation carriers (BOSOM), which is a Dutch consecutive series of 7,106 patients with invasive BC treated for their primary BC in ten centers throughout the Netherlands between 1970 and 2003; in this study 94% of patients were genotyped for *BRCA1/2* germline mutations[3]; (4) The Erasmus Medical Center (EMC) study containing patients diagnosed with BC between 1989 and 2013 who were treated at the EMC in Rotterdam; for this study, complete follow-up was obtained for 3,483 patients that had been diagnosed between 2000 and 2009; (5) The Netherlands Cancer Registry (NCR), which is an ongoing nationwide population-based data registry of all newly diagnosed cancer patients in the Netherlands since 1989[4]. We included patients diagnosed between 2003 and 2010, a period for which sufficient follow-up and receptor status information were provided[4, 5]. The eligibility criteria applied in each data source is reported in **Table S1**. Data were harmonized by recoding each of the main datasets by the responsible data managers according to a standardized data dictionary. We performed checks for data consistency and validity centrally.

We extracted the following information: *BRCA1/2* germline mutation, family history (first degree) of primary BC, and regarding primary BC diagnosis: age, nodal status, size, grade, morphology, estrogen-receptor (ER) status, progesterone-receptor (PR), human epidermal growth factor receptor 2 (HER2) status, administration of adjuvant or neoadjuvant chemotherapy, adjuvant endocrine therapy, adjuvant trastuzumab therapy, radiotherapy. We excluded PR status and TNM stage of the primary BC due to collinearity with ER status and the size of the primary tumor, respectively. In the current clinical practice, only patients with ER-positive tumors receive endocrine therapy and only patients with HER2-positive tumors receive trastuzumab; these co-occurrences were considered in the model by using composite categorical variables. A description of the studies included in the analyses is provided in **Table S2**. Follow-up started three months after invasive first primary BC diagnosis, in order to exclude synchronous CBCs, and ended at date of CBC, distant metastasis (but not at loco-regional relapse), CPM, or last date of follow-up (due to death, being lost to follow-up, or end of study), whichever

occurred first. We considered that after loco-regional relapse, a woman would be still at risk for CBC as treatment for loco-regional relapse would not affect the contralateral breast cancer (CBC) unless adjuvant systemic treatment was given. Distant metastasis was considered as a competing risk because most of the patients receive systemic therapies after developing distant metastasis.

Table S1. Data source flowchart.

	Source of data				
	ABCS	**BCAC‡**	**BOSOM**	**EMC**	**NCR**
Number of patients	2,390	182,898	7,105	3,483	94,600
Eligibility criteria, *number of patients excluded*					
Studies from Asian countries	-	7,348	-	-	-
Patients of non-European descent	-	46,670	-	-	-
Year of PBC diagnosis before 1990	-	3,358	3,126	-	-
Year of PBC diagnosis missing	-	26,291	-	-	-
PBC stage 0	122	34	2	-	-
PBC stage IV	94	1,675	104	-	4,569
Patients did not undergo surgery	24	1,138	43	5	5,174
Number of eligible patients	2,150	96,384	3,830	3,478	84,857
No follow-up or follow-up less than 3 months	171	13,144	70	88	1,719
Familiar breast cancer studies	-	4,635	-	-	-
Studies with less than 10 CBC events	-	38,116	-	-	-
Number of patients included in the analysis (number of patients with CBC)	1,979 (19)	40,489 (707)	3,760 (288)	3,390 (221)	83,138 (3,447)
Total number of patients included in the analysis (number of CBC)	132,756 (4,682 of which 3,974 invasive and 708 *in-situ*)				

Abbreviations: ABCS: Amsterdam Breast Cancer Study; BCAC: Breast Cancer Association Consortium. ‡BCAC is composed of 102 studies world-wide. The 40,489 patients selected for the analysis came from 16 studies.; BOSOM: Breast Cancer Outcome Study of Mutation carriers; EMC: Erasmus Medical Center; NCR: Netherlands Cancer Registry; PBC: primary breast cancer; CBC: contralateral breast cancer

Table S2: see online material

Age at first primary BC seemed to have a non-linear relationship with CBC; using splines we observed that CBC risk increased with age till around 50 years old and declined afterwards; see **Figure S1**. Therefore, we used a linear spline with a knot at 50 years in the prediction model. The use of this linear spline was a good compromise to address the non-linear relationship between CBC risk and age across the different baseline risks in all the studies, with different age distributions and selections (one study included only women aged under 50 years). Moreover, the observed non-linear relationship resembled the shape of age-related BC incidence curves with an increased risk until menopausal age followed by a decrease (Clemmensen's hook)[6].

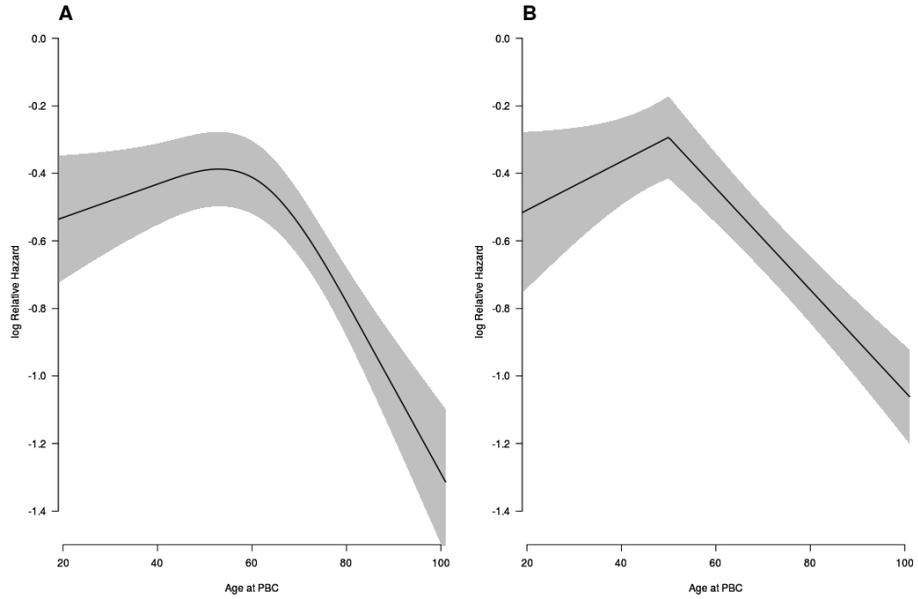


Figure S1: Graphical assessment of non-linear relationship of age with contralateral breast cancer risk.
A non-linear relationship between age at first primary breast cancer (x-axis) and the log relative hazard of contralateral breast cancer (y-axis) is shown. Panel A shows a restricted cubic spline with three knots. Panel B shows a linear spline with one knot located at 50 years. The curve gray area indicates the corresponding 95% confidence intervals. Both curves were estimated from a multivariable subdistributional hazard model adjusted for the variables used for the risk prediction considering death for any causes and distant metastasis as a competing risk.

2. Multiple imputation of missing values

The percentage of missing values across the predictors varied between 5.1% and 94.2% for morphology of first primary BC and *BRCA* mutation, respectively. In the individual patient data (IPD), both sporadic and systematic missing may occur. The former are missing values within a study, the latter are values missing for all individuals within a particular study[7-9].

For our analyses, we used ten imputed datasets based on the multiple imputation chained equations (MICE) using 50 iterations. The visit sequence of the variables was in ascending order of the number of missing values. This technique improves the accuracy and the statistical power assuming missing is at random (MAR). In the imputation procedure, we also used the year of first primary BC diagnosis since this information provides a better correlation structure among covariates used as predictors in the imputation model. Since there were systematic missing data, we used the imputation model based on the stratified multiple imputation strategy (SMI). In this approach, the variable identifying

the study was used as covariate to improve substantially the imputation especially for the systematic missing predictors that might occur in the individual patient data (IPD) from multiple studies[9]. Continuous, binary and multiple categorical variables were imputed using predictive mean matching, binary and polytomous logistic regression, respectively. Time-to-event outcome defined as time to contralateral breast, time to death, and time to distant metastasis were included in the imputation process through the Nelson-Aalen cumulative hazard estimator[10]. For every variable with missing data, every imputation model selects predictors based on correlation structure underlying the data. We recoded the variables chemotherapy and morphology after imputation. In particular, information about neoadjuvant and adjuvant chemotherapy were separately imputed. Then, we created a chemotherapy variable by combining the variables for neoadjuvant and adjuvant chemotherapy in every imputed dataset. Morphology of primary tumor was imputed by keeping all original categories ('Lobular', 'Ductal', 'Mixed' and 'Other'). After multiple imputation, we created two categories 'Lobular including mixed' and 'Ductal including other' to address possible overfitting due to the small samples of 'Mixed' and 'Other' categories. Since in current clinical practice, only estrogen receptor (ER) positive patients receive endocrine therapy and only human epidermal growth factor receptor 2 (HER2) positive patients receive trastuzumab, composite categorical factors of ER and endocrine therapy and of HER2 and trastuzumab therapy were considered in the model building. However, in our data, 2% of patients with 70 CBC events were coded as ER-negative treated with endocrine therapy and 0.2% of patients with 7 CBC events were coded as HER2-negative treated with trastuzumab therapy. In every imputed dataset, we recoded those patients as ER-positive treated with endocrine treatment and HER2-positive treated with trastuzumab since the largest proportion of patients (53%) were ER-positive treated with endocrine therapy and 82% were HER2-positive treated with trastuzumab in the complete data.

We used the R package mice (version 2.46.0) to impute our data and combine the estimates using Rubin's rules.

3. Complete case analysis

When a missing data pattern is completely at random (MCAR), imputation of missing data is not necessary. Therefore, descriptive analyses were performed to check whether the missing data pattern was MCAR. For completeness, the patients and first primary breast cancer characteristics and results of the multivariable subdistributional hazard model based on the case set with complete data are shown in **Table S3** and **Table S4,** respectively. The prediction performance of the risk prediction model was not investigated since in the case set with complete data all cases came from one geographic area (Western Europe) and the number of CBC event did not reach the number of events required for an external validation.

Table S3: Patients and first primary breast cancer characteristics used in the contralateral breast cancer risk prediction model in the complete case and all case analyses.

Factors at primary breast cancer		N	%
		132,756	**100.0**
Age, *years*			
	Median (range)	57 (18 - 101)	
	Missing	-	
Family history			
	Yes	5,959	19.5
	No	24,582	80.5
	Missing	102,215	-
BRCA mutation			
	BRCA1	333	4.3
	BRCA2	167	2.2
	Non carrier	7,204	93.5
	Missing	125,052	-
Nodal status			
	Positive	48,979	39.1
	Negative	76,356	60.9
	Missing	7,421	-
Tumor size, *cm*			
	≤ 2	75,849	60.8
	(2-5]	43,075	34.5
	> 5	5,916	4.7
	Missing	7,916	-
Tumor grade			
	well differentiated	25,271	21.7
	moderately differentiated	53,385	45.7
	poorly/undifferentiated	38,045	32.6
	Missing	16,055	-
ER status			
	Positive	97,460	80.5
	Negative	23,625	19.5
	Missing	11,671	-
HER2 status			
	Positive	15,401	17.4
	Negative	72,891	82.6
	Missing	44,464	-
Morphology			
	Ductal	96,561	76.6
	Lobular	14,681	11.7
	Mixed	4,982	4.0
	Other	9,780	7.8
	Missing	6,752	-
Adjuvant chemotherapy			
	Yes	46,868	38.2
	No	75,785	61.8
	Missing	10,103	-
Neoadjuvant chemotherapy			
	Yes	7,213	6.0
	No	112,267	94.0
	Missing	13,276	-

Table S3: Continued

Factors at primary breast cancer		N	%
		132,756	100.0
Endocrine adjuvant therapy			
	Yes	65,959	54.1
	No	56,055	45.9
	Missing	10,742	-
Trastuzumab adjuvant therapy			
	Yes	6,875	6.7
	No	9,6324	93.3
	Missing	29,557	-
Radiation in the breast			
	Yes	85,029	69.5
	No	37,237	30.5
	Missing	10,490	-
CBC cumulative incidence, *%*			
	5-year (95%CI)	2.1 (2.1 - 2.2)	
	10-year (95%CI)	4.1 (4.0 - 4.3)	

Abbreviations:
PBC: primary breast cancer; ER: estrogen-receptor; HER2: human epidermal growth factor receptor 2; CBC: contralateral breast cancer; CI: confidence interval.

Table S4: Results of multivariable subdistributional hazard model using the complete case dataset.

Factor (categories) at primary breast cancer	Multivariable analysis	
	sHR	95% CI
Age, *years*	1.48[a]	0.73 – 3.00[a]
Family history (yes versus no)	1.36	0.69 - 2.70
BRCA mutation		
BRCA1 versus non-carrier	5.28	2.13 - 13.10
BRCA2 versus non-carrier	2.30	0.50 - 10.51
Nodal status (positive versus negative)	1.37	0.56 - 3.34
Tumor size, *cm*		
(2,5] versus ≤ 2	0.57	0.22 - 1.47
> 5 versus ≤ 2	3.53	1.10 - 11.34
Morphology (lobular including mixed versus ductal including other)	0.99	0.33 - 2.88
Grade		
Moderately differentiated versus well differentiated	0.91	0.28 - 2.88
Poorly differentiated versus well differentiated	0.84	0.23 - 3.04
Chemotherapy (yes versus no)	0.38	0.16 – 0.89
Radiotherapy to the breast (yes versus no)	1.26	0.56 - 2.83
ER (positive or negative) / endocrine therapy (yes or no)		
negative/no versus positive/yes	1.42	0.53 - 3.77
positive/no versus positive/yes	2.38	0.90 - 6.31
HER2 (positive or negative) / trastuzumab therapy (yes or no)		
negative/no versus positive/yes	0.71	0.22 - 2.36
positive/no versus positive/yes	0.32	0.07 - 1.46

Abbreviations:
sHR: subdistributional hazard ratio; CI: confidence interval; ER: estrogen receptor; HER2: human epidermal growth factor receptor 2;
[a] Age was parameterized as a linear spline at 50. For presentation purposes, we here provide the sHR for the 75th versus the 25th percentile.

4. Model diagnostics and baseline recalibration

For the multivariable model, we checked the assumption of proportional subdistribution hazards graphically using Schoenfield residuals. Heterogeneity of baseline risks between studies was taken into account using the study as a stratification term. We estimated the cumulative incidence at 5- and 10- year using the baseline hazard of the Netherlands Cancer Registry (NCR) dataset to improve the model calibration, since this is our largest cohort (67% of the data) and is based on complete incidence data thus provides a representative cumulative incidence of CBC (4.6% at 10 years). The stratified model and the application of the Rubin's rules took into account both the between study and between imputation variation.

5. Leave-one-study-out cross-validation

We used leave-one-study-out cross-validation (also known as an internal-external validation), in which a model for predicting CBC risk is developed in all studies except one whose external validity is evaluated (every study is excluded once in this process). For the studies where the number of CBC events was insufficient for external validation, we used the geographic area as a unit of splitting. For time-to-event outcomes at least 100 events per study are required for external validation[11]. The geographic area corresponding to every study is shown in **Table S5**.

Table S5: List of BCAC studies (including ABCS source) with the corresponding country and geographic area. For studies in which the number of contralateral breast cancer events was insufficient for external validation, the geographic area was used.

Study	Country	Geographic area or study
ABCS	Netherlands	Europe - Other
ABCFS	Australia	United States and Australia
BBCC	Germany	Europe - Other
CGPS	Denmark	Europe - Scandinavia
HEBCS	Finland	Europe - Scandinavia
KARBAC	Sweden	Europe - Scandinavia
KARMA	Sweden	Europe - Scandinavia
LMBC	Belgium	Europe - Other
MARIE	Germany	Europe - Other
MEC	United States	United States and Australia
ORIGO	Netherlands	Europe - Other
PBCS	Poland	Europe - Other
PKARMA	Sweden	Europe - Scandinavia
POSH	United Kingdom	Europe - United Kingdom
SEARCH	United Kingdom	Europe - United Kingdom
SKKDKFZS	Germany	Europe - Other
SZBCS	Poland	Europe - Other

Table S6: see online material

We evaluated the discrimination accuracy using the time-dependent area under the curve (AUC) at 5- and 10-year. The Inverse Probability of Censoring Weighting (IPCW) was computed to estimate of cumulative/dynamic time-dependent AUCs[12]. Since the mortality and distant metastasis were competing risks, a control was defined as a subject not experiencing a CBC at 5- and 10-year, respectively. The AUC estimate and the corresponding confidence intervals were computed by bootstrapping 100 times every imputed dataset in each validation study. The AUCs and the corresponding confidence intervals were pooled using Rubin's rules.

We did not consider delayed-entry patients (with prevalent BC) to evaluate the discrimination accuracy of the prediction models since no standard performance measures are currently available in the statistical literature to account for left-truncated follow-up time. In our study the median of delayed entry was 0.6 years.

We assessed the calibration of the models using calibration-in-the-large, calibration slope, and calibration plots per study[13]. Calibration plots report the predicted probabilities on the x-axis and the observed probabilities on the y-axis. For time-to-event data, this plot can be generated at multiple time points. To reproduce the nomogram building, we used the predicted and observed cumulative incidence of 5- and 10-year as time points for the calibration plots. The observed and predicted outcomes are divided by quartiles of predicted values. In case of good overall calibration, all points in a calibration plot are near the 45-degree line starting at the origin (0,0). If points are below the 45-degree line, models overestimate the observed risk (overfitting). If points are above the 45-degree line, the model underestimates the observed risk (underfitting). In each validation study, calibration slopes and predicted probabilities at 5 and 10 years were calculated in every imputed dataset. Then, for each validation study, calibration slopes and the predicted probabilities were pooled using Rubin's rules. Calibration plots at 5- and 10-year are shown in **Figures S2 and S3,** respectively.

6. Clinical utility

The decision curve analysis combines the direct applicability of the decision-analytic methods with the mathematical simplicity of accuracy metrics[14]. The mathematical background of the net benefit calculation was originally developed by Peirce in 1884[15]. More recently, other publications expanded this work and proposed and gave emphasis why the net benefit measures should be used beyond measures of discrimination and calibration to assess the accuracy of prediction models[16].
The net benefit (NB) is calculated as:

$$NB = \frac{TP}{n} - \frac{FP}{n}\left(\frac{p_t}{1-p_t}\right)$$

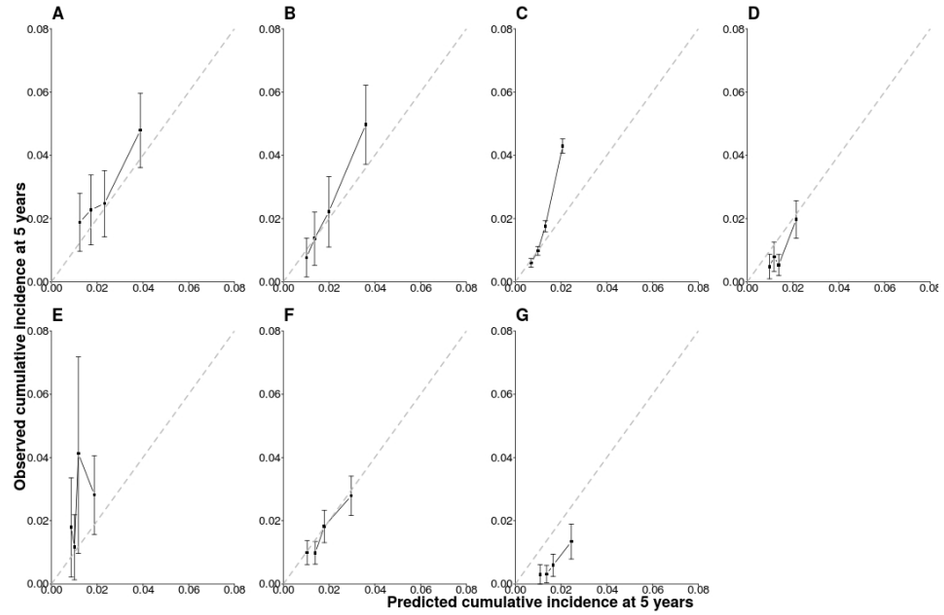


Figure S2: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 5 years for the contralateral breast cancer risk model with *BRCA* mutation information. The x-axis represents the predicted cumulative incidence of contralateral breast cancer at 5 years and the y-axis the observed cumulative incidence at 5 years. The black dots indicate the calibration for quartiles of predicted values. Vertical black bars indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Netherlands - BOSOM; Panel B: Netherlands - EMC; Panel C: Netherlands - NCR; Panel D: Europe – Scandinavia; Panel E: United States and Australia; Panel F: Europe – Other; Panel G: Europe – United Kingdom.

Where n is the total sample size *TP* = true positive counts; *FP* = false positive counts; p_t = risk threshold that defines the high risk and low risk patients. The ratio $\frac{p_t}{1-p_t}$ represents the relative weight of the harm of unnecessary contralateral preventive mastectomies (CPM) versus the benefit of CBC patients who truly need the surgery. To draw the decision curve, the net benefit is calculated for different values of p_t.

The risk thresholds and the calculation of the true positives and false negatives in case of censored data with competing risks are defined as:

$$TP = \{I(t)|X = 1\} \cdot P(X = 1) \cdot n$$
$$FP = \{1 - I(t)|X = 1\} \cdot P(X = 1) \cdot n$$

Where:
n = total number of BC patients;
I(t) = cumulative incidence of CBC predicted by the prediction model at time *t*;

$$X = \begin{cases} 1\ predicted\ \ cumulative\ incidence\ at\ time\ t\ \geq\ p_t \\ 0\ predicted\ \ cumulative\ incidenceat\ time\ t\ <\ p_t \end{cases}$$

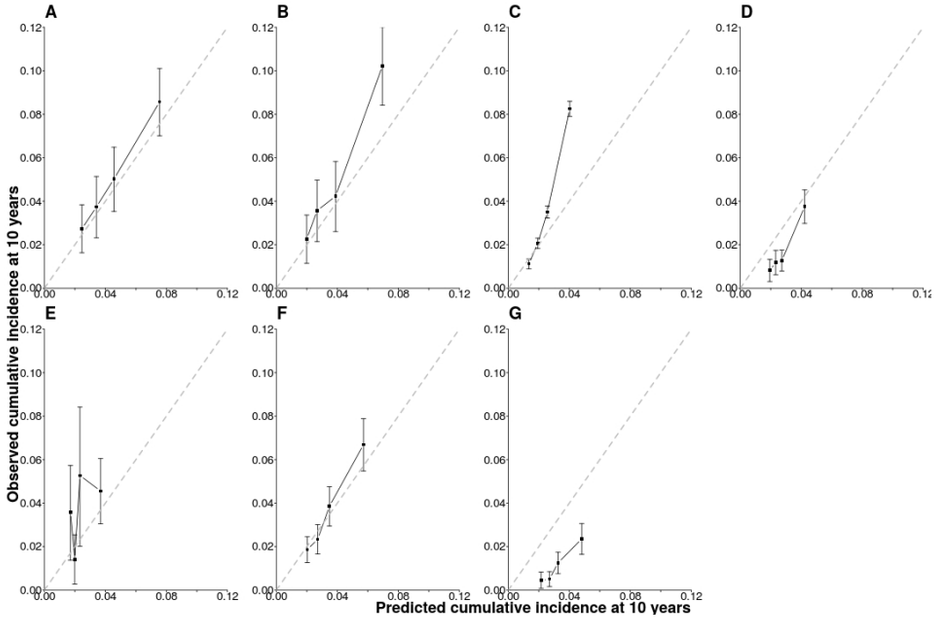


Figure S3: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 10 years for the contralateral breast cancer risk model with *BRCA* mutation information. The x-axis represents the predicted cumulative incidence of contralateral breast cancer at 10 years and the y-axis the observed cumulative incidence at 10 years. The black dots indicate the calibration for quartiles of predicted values. Vertical black bars indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Netherlands - BOSOM; Panel B: Netherlands - EMC; Panel C: Netherlands - NCR; Panel D: Europe – Scandinavia; Panel E: United States and Australia; Panel F: Europe – Other; Panel G: Europe – United Kingdom.

More mathematical details have been provided by Vickers in 2008 and Kerr in 2016[17, 18]. The landmark time *t* was set to 5 and 10 years since the prediction model provided the estimated cumulative incidence at 5 and 10 years.

Although discrimination measures such as sensitivity, specificity, Area Under the Curve (AUC), and c-statistic and calibration measures cannot be used to assess the clinical utility of a prediction model, net benefit is larger for more discriminating models and decrease with poor calibration[19]. Referring to our model, the reduction in the number of unnecessary CPM per 1,000 patients without a decrease in the number of patients who correctly received the surgery is calculated as:

$$(net\ benefit\ of\ the\ model - net\ benefit\ of\ treat\ all)/(\ pt\ /(1 - pt))\ \times\ 1{,}000$$

For example, at a risk threshold of 10% the difference between the net benefit of the prediction model and the net benefit of treat all was 0.0179, the number of avoidable unnecessary CPM would be [0.0179/(0.10/0.90)] = 0.1611 x 1000 = 161.1 per 1,000 patients.

Results of the decision curve analysis that were not reported in Table 2, are reported in **Table S7.** The utilization of 5-year CBC risk prediction in terms of net benefit showed that for some risk thresholds (between 1.5–4.5%), the prediction model might be clinically useful to avoid unnecessary CPM among *BRCA1* patients and to counsel necessary CPM among non-carriers. As an example, if a clinician finds it acceptable to perform around 21 unnecessary CPM to prevent one CBC (one necessary CPM), a risk threshold of 4.5% may be used to define high and low risk *BRCA1/2* patients based on the absolute 5-year CBC risk prediction estimated by the model. In this scenario, approximately 163 CPMs per 1,000 patients may be avoided using the model compared to counseling CPM to all *BRCA1/2* carriers. Similarly, if unnecessarily performing a CPM in 39 patients would be acceptable to prevent one CBC, a risk threshold of 2.5% may be used to define high and low risk non-carriers; and this would include around necessary 491 CPMs per 1,000 patients. The decision curves in **Figures S4** provide a comprehensive overview of the net benefit for a range of harm-benefit thresholds at 5-year CBC risk.

Table S7: Clinical utility of the 5-year contralateral breast cancer risk prediction model. At the same probability threshold, the net benefit is exemplified in *BRCA1/2* mutation carriers (for avoiding unnecessary CPM) and non-carriers (performing necessary CPM).

		BRCA1/2 mutation carriers		Non-carriers	
Probability threshold p_t (%)	Unnecessary CPMs needed to prevent a CBC*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients
1.5	65.7	0.0	0.0	3.3	216.7
2.5	39.0	0.1	3.9	12.6	491.4
3.5	27.6	2.3	63.4	0.1	2.8
4.5	21.2	7.7	163.4	0.0	0.0

CPM: contralateral preventive mastectomy; CBC: contralateral breast cancer;
* The number of unnecessary contralateral mastectomies needed to prevent a CBCis calculated by: $(1\text{-}p_t)/p_t$

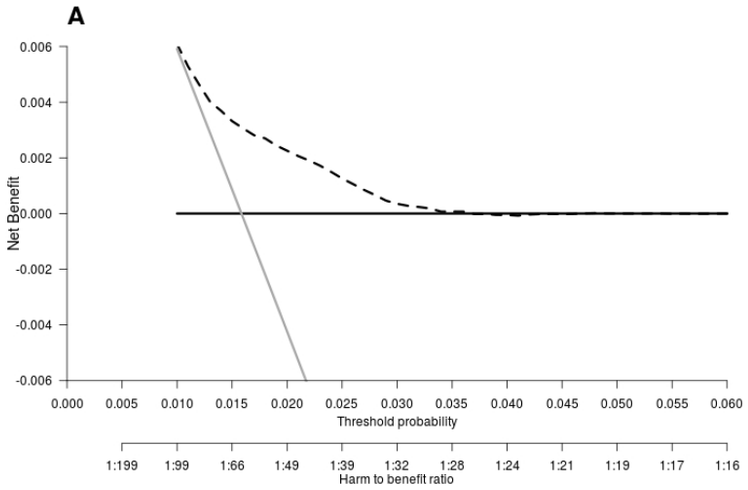


Figure S4: Decision curve analysis at 5 years for the contralateral breast cancer risk model including *BRCA1/2* mutation information.
Panel A shows the decision curve to determine the net benefit of the estimated 5-year predicted contralateral breast cancer (CBC) cumulative incidence for patients without a *BRCA1/2* gene mutation using the prediction model (dotted black line) compared to not treating any patients with contralateral preventive mastectomy (CPM) (black solid line). Panel B shows the decision curve to determine the net benefit of the estimated 5-year predicted CBC cumulative incidence for *BRCA1/2* mutation carriers using the prediction model (dotted black line) versus treating (or at least counseling) all patients (grey solid line). The y-axis measures net benefit, which is calculated by summing the benefits (true positives, i.e., patients with a CBC who needed a CPM) and subtracting the harms (false positives, i.e., patients with CPM who do not need it). The latter are weighted by a factor related to the relative harm of a non-prevented CBC versus an unnecessary CPM. The factor is derived from the threshold probability to develop a CBC at 5 years at which a patient would opt for CPM (e.g. 4.5%). The x-axis represents the threshold probability. Using a threshold probability of 4.5% implicitly means that CPM in 22 patients of whom one would develop a CBC if untreated is acceptable (21 unnecessary CPMs, harm to benefit ratio 1:21).

7. Formula to estimate the contralateral breast cancer risk

Our developed model is a subdistributional proportional hazard Fine and Gray model. The estimated cumulative incidence of CBC was estimated using the following formula:

$$F(t) = 1 - \left\{[S_0(t)]^{\exp(\boldsymbol{LP})}\right\}$$

Where *t* is the time (in years) since primary BC, *F* (*t*) is the cumulative incidence of CBC and $S_0(t)$ is the probability to survive beyond for baseline covariate values. The baseline survival estimates according to the model and time are:

$S_0(5) = 0.984$
$S_0(10) = 0.968$

And

Linear Predictor (LP) =

– 0.223 + 0.007× Age – 0.023 × Age' + 0.303 × I[Family history = Yes] + 1.304 × I[*BRCA* = *BRCA1*] + 0.941 × I[*BRCA* = *BRCA2*] – 0.142 × I[Nodal status = positive] – 0.047 × I[Size of PBC = (2,5] *cm*] + 0.128 × I[Size of PBC = greater than 5 *cm*] + 0.209 × I[Morphology of PBC = lobular including mixed] – 0.120× I[Grade of PBC = moderately differentiated] – 0.291 × I[Grade of PBC = poorly/undifferentiated]– 0.266 × I[Chemotherapy = yes] + 0.009 × I[Radiotherapy to the breast = yes] + 0.356 × I[ER-negative without endocrine therapy] + 0.559 × I[ER-positive without endocrine therapy]+ 0.082 × I[HER2-negative without trastuzumab] – 0.005 × I[HER2-positive without trastuzumab]

Where Age' = max(Age – 50, 0)

8. Results of the prediction model without BRCA mutation

Because a patient may not have been tested for the *BRCA* gene mutations, this information may not be available before or on the day of first primary BC diagnosis or treatment decisions. Moreover, information about *BRCA* mutations was largely missing in the databases we used. Thus, we also developed and validated a CBC prediction model without *BRCA* mutations to also provide an individualized risk prediction tool for patients not tested. Results of the risk prediction model in terms of relative subdistributional hazard ratio (sHRSs) and the corresponding 95% confidence intervals (CI) for patients not tested for *BRCA* gene mutations are provided in **Table S8**.

The assessments of prediction performance are shown in **Figures S5, S6, and S7.** The discrimination accuracy at 5 years was 0.59 (95% CI: 0.54 – 0.63; 95% prediction interval (PI): 0.46 – 0.71) and at 10 years was 0.59 (95% CI: 0.56 – 0.62; 95% PI: 0.52 – 0.66), as shown in **Figure S5**. The calibration-in-the-large was -0.17 (95% CI: -0.72 – 0.38; 95% PI: -1.70– 1.36), as shown in the **Figure S5 panel C;** and calibration slope was 0.81 (95% CI: 0.63 – 0.99; 95% PI: 0.50 – 1.12) in the leave-one-study-out cross-validation, as shown

in **Figure S5 panel D**. The calibration plots at 5- and 10-year are reported in **Figures S6 and S7**, respectively.

Table S8: Results of multivariable subdistributional hazard model for breast cancer patients without *BRCA* mutations.

Factor (category) at primary breast cancer	Multivariable analysis	
	sHR	95% CI
Age, *years*	0.61[a]	0.56-0.66[a]
Family history (yes versus no)	1.60	1.50 - 1.71
Nodal status(positive versus negative)	0.87	0.80 - 0.93
Tumor size, *cm*		
(2,5] versus ≤ 2	0.96	0.89 - 1.03
> 5 versus ≤ 2	1.11	0.97 - 1.28
Morphology (lobular including mixed versus ductal including other)	1.20	1.10 - 1.30
Grade		
Moderately differentiated versus well differentiated	0.97	0.90 - 1.04
Poorly differentiated versus well differentiated	0.87	0.79 - 0.96
Chemotherapy (yes versus no)	0.78	0.71 - 0.85
Radiation of the breast (yes versus no)	0.97	0.90 - 1.03
ER (positive or negative) / endocrine therapy (yes or no)		
negative/no versus positive/yes	1.67	1.54 - 1.84
positive/no versus positive/yes	1.81	1.67 - 1.96
HER2 (positive or negative) / trastuzumab therapy (yes or no)		
negative/no versus positive/yes	1.26	1.08 - 1.48
positive/no versus positive/yes	1.08	0.91 - 1.30

Abbreviations:
sHR: subdistributional hazard ratio; CI: confidence interval; ER: estrogen receptor; HER2: human epidermal growth factor receptor 2;
[a]: Age was parameterized as a linear spline at 50. For representation purposes, we here provide the sHR for the 75th versus the 25th percentile.

Table S9. Clinical utility of the 5-year contralateral breast cancer risk prediction model in non-*BRCA* tested patients. At the same probability threshold, the net benefit is exemplified in patients with family history (for avoiding unnecessary CPM) and patients without family history (performing necessary CPM).

		Family history		No family history	
Probability threshold p_t (%)	Unnecessary CPMs needed to prevent a CBC*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients
2.0	49.0	0.4	19.6	2.1	102.9
2.5	39.0	2.9	113.1	1.2	46.8
3.0	32.3	0.0	0.0	0.2	6.5

CPM: contralateral preventive mastectomy; CBC: contralateral breast cancer;
* The number of unnecessary contralateral preventive mastectomies needed to prevent a CBC is calculated by: $(1-p_t)/p_t$

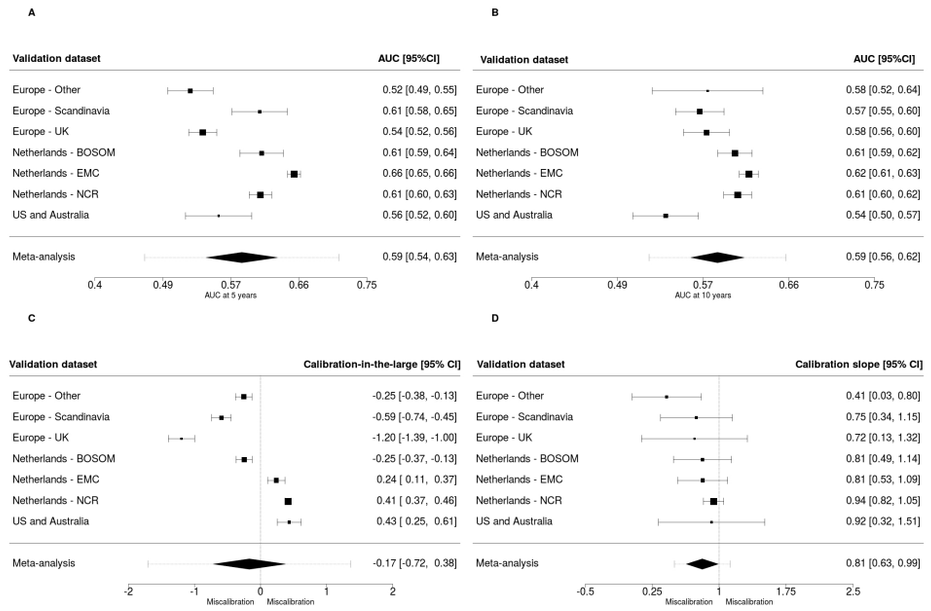


Figure S5: Results of the leave-one-study-out cross-validation for the contralateral breast cancer risk model at 5 and 10 years without *BRCA* mutation information.
Panel A and B show the discrimination accuracy assessed by a time-dependent AUC at 5 and 10 years, respectively. Panel C shows the calibration accuracy measured with calibration in-the-large. Panel D shows the calibration accuracy measured with calibration slope. The black squares indicate the estimated accuracy of the model in a single new validation study or geographic area. The black horizontal lines interval indicate the corresponding 95% confidence intervals of the estimated accuracy (interval whiskers). The black diamonds indicate the mean with the corresponding 95% confidence interval of the predictive accuracy and the dashed horizontal lines indicate the corresponding 95% prediction intervals.

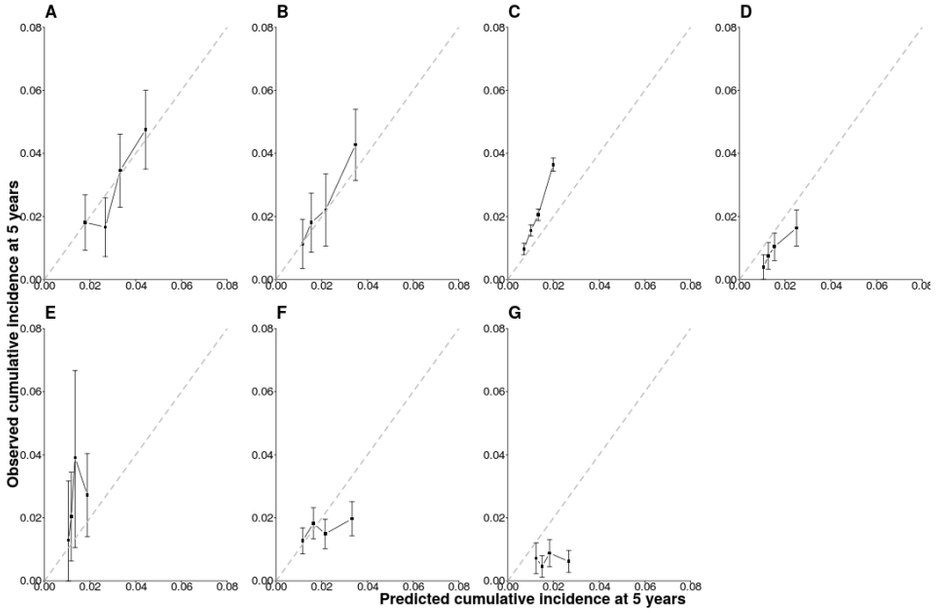


Figure S6: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 5 years for the contralateral breast cancer risk model without *BRCA* gene mutation information.

The x-axis represents the predicted cumulative incidence of contralateral breast cancer at 5 years and the y-axis the observed cumulative incidence at 5 years. The black dots indicate the calibration for quartiles of predicted values. Vertical black bars indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Netherlands - BOSOM; Panel B: Netherlands - EMC; Panel C: Netherlands - NCR; Panel D: Europe – Scandinavia; Panel E: United States and Australia; Panel F: Europe – Other; Panel G: Europe – United Kingdom.

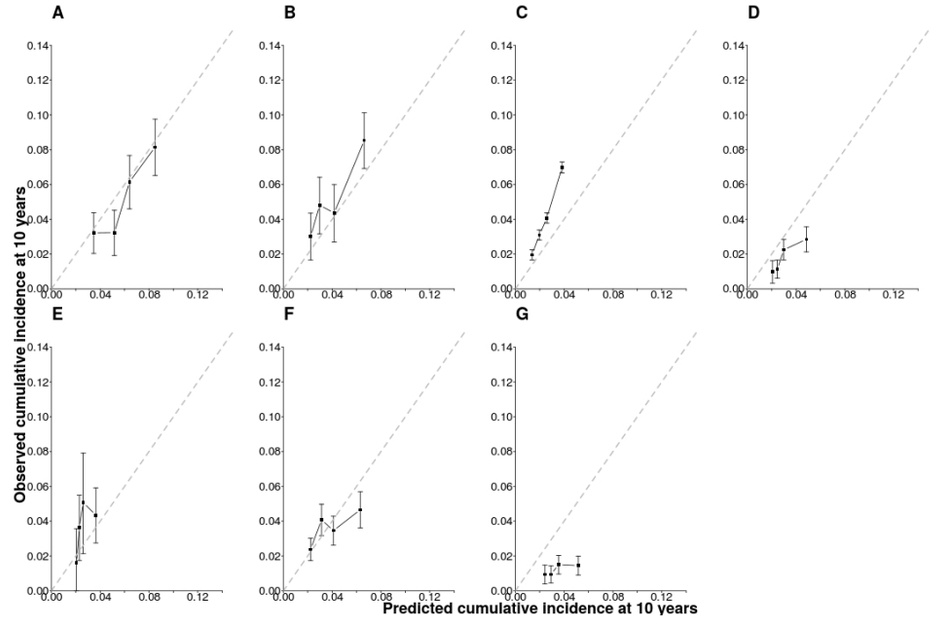


Figure S7: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 10 years for the contralateral breast cancer risk model without *BRCA* gene mutation information.

The x-axis represents the predicted cumulative incidence of contralateral breast cancer at 10 years and the y-axis the observed cumulative incidence at 10 years. The black dots indicate the calibration for quartiles of predicted values. Vertical black bars indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Netherlands - BOSOM; Panel B: Netherlands - EMC; Panel C: Netherlands - NCR; Panel D: Europe – Scandinavia; Panel E: United States and Australia; Panel F: Europe – Other; Panel G: Europe – United Kingdom.

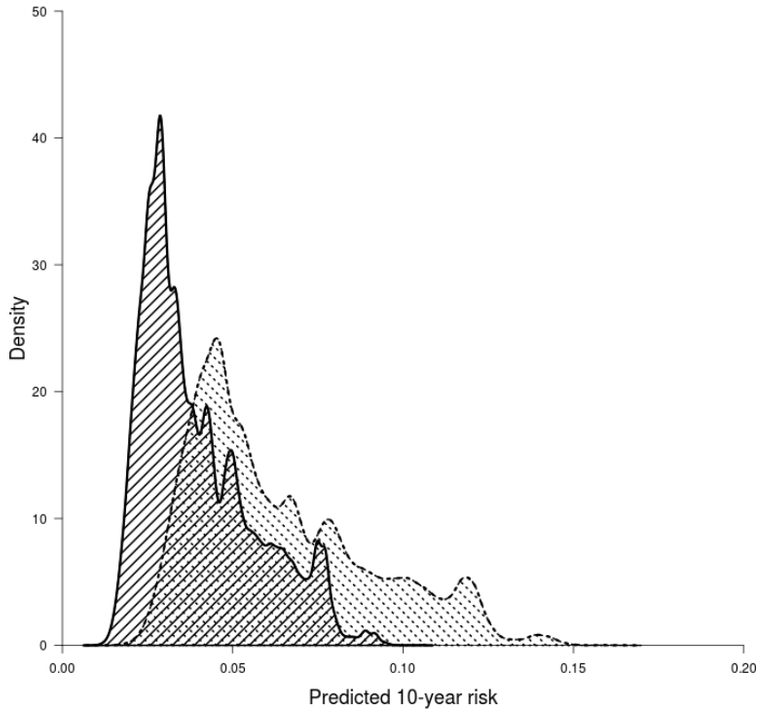


Figure S8: Density distribution of 10-year predicted absolute risk in patients with no family history (area with black lines) and patients with a family history (area with dashed lines).

The utilization of 10-year CBC risk prediction in terms of net benefit showed that for some risk thresholds (between 3.5–5.5%), the prediction model might be clinically useful to avoid unnecessary CPM among patients with family history and to counsel necessary CPM among patients without first-degree relatives with BC. For example, if a clinician finds it acceptable to perform around 21 unnecessary CPM to prevent one CBC, a risk threshold of 4.5% may be used to define high and low risk patients with family history based on the absolute 10-year CBC risk prediction estimated by the model. In this scenario, 55 CPM per 1,000 patients may be avoided using the model compared to counseling CPM to all patients with family history; see **Table S10** and **Figure S8**. The density distribution of the estimated 10-year CBC risk prediction was shown for patients with and without family history. The overlap between the two distributions reflects that a prediction model is useful to define high and low risk patients to counsel necessary CPM and avoid unnecessary surgeries in a setting where *BRCA1/2* mutations are not tested. The decision curves in **Figures S9 and S10** provide a comprehensive overview of the net benefit for a range of harm-benefit thresholds.

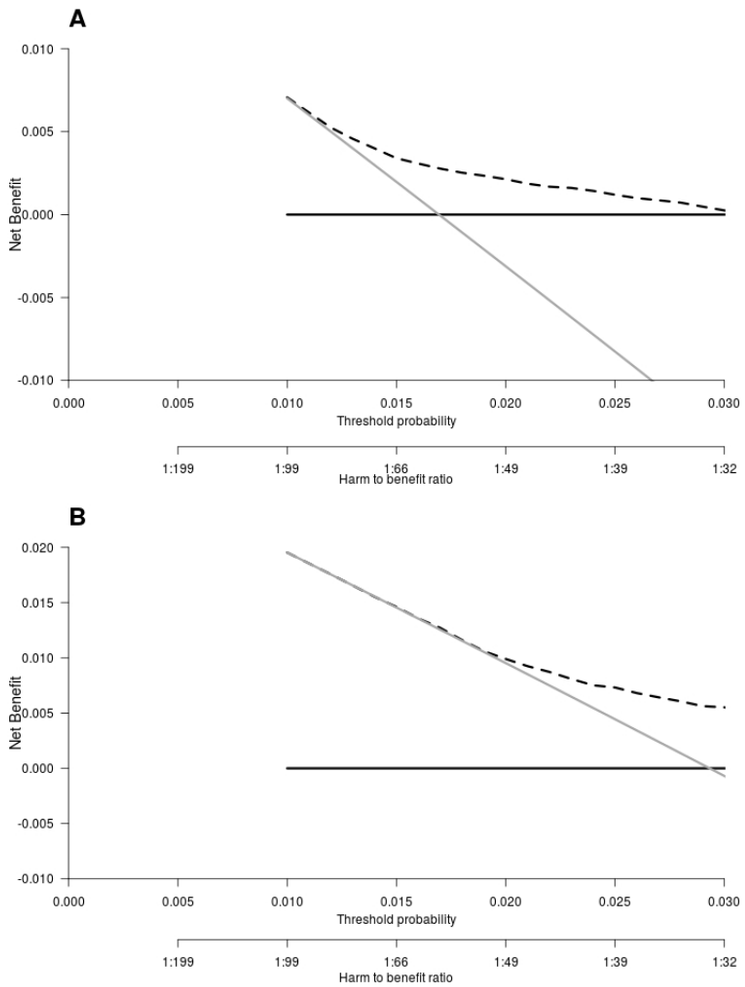


Figure S9: Decision curve analysis at 5 years for the contralateral breast cancer risk model without *BRCA* mutation information.
Panel A shows the decision curve to determine the net benefit of the estimated 5-year predicted contralateral breast cancer (CBC) cumulative incidence for patients without first-degree family history using the prediction model (dotted black line) compared to not treating any patients with contralateral preventive mastectomy (CPM) (black solid line). Panel B shows the decision curve to determine the net benefit of the estimated 5-year predicted CBC cumulative incidence for patients with first-degree family history of breast cancer using the prediction model (dotted black line) versus treating (or at least counseling) all patients (grey solid line). The y-axis measures net benefit, which is calculated by summing the benefits (true positives, i.e., patients with a CBC who needed a CPM) and subtracting the harms (false positives, i.e., patients with CPM who do not need it). The latter are weighted by a factor related to the relative harm of a non-prevented CBC versus an unnecessary CPM. The factor is derived from the threshold probability to develop a CBC at 5 years at which a patient would opt for CPM (e.g. 2.5%). The x-axis represents the threshold probability. Using a threshold probability of 2.5% implicitly means that CPM in 40 patients of whom one would develop a CBC if untreated is acceptable (39 unnecessary CPMs, harm to benefit ratio 1:39).

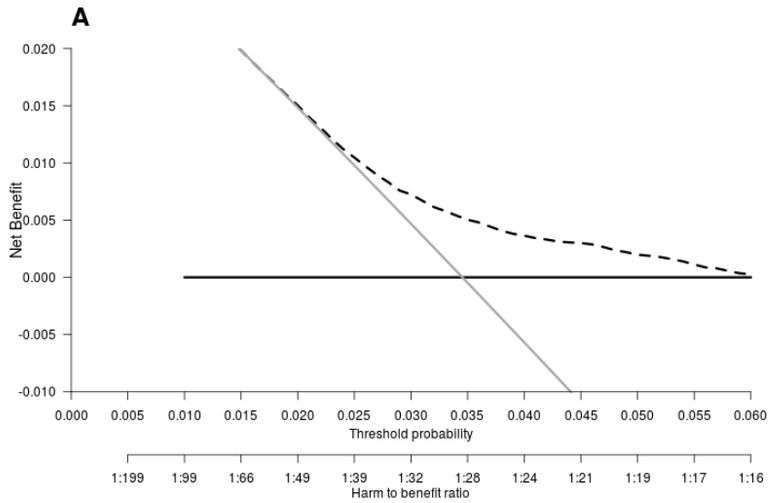


Figure S10: Decision curve analysis at 10 years for the contralateral breast cancer risk model without *BRCA* mutation information.
Panel A shows the decision curve to determine the net benefit of the estimated 10-year predicted contralateral breast cancer (CBC) cumulative incidence for patients without first-degree family history using the prediction model (dotted black line) compared to not treating any patients with contralateral preventive mastectomy (CPM) (black solid line). Panel B shows the decision curve to determine the net benefit of the estimated 10-year predicted CBC cumulative incidence for patients with first-degree family history using the prediction model (dotted black line) versus treating (or at least counseling) all patients (grey solid line). The y-axis measures net benefit, which is calculated by summing the benefits (true positives, i.e., patients with a CBC who needed a CPM) and subtracting the harms (false positives, i.e., patients with CPM who do not need it). The latter are weighted by a factor related to the relative harm of a non-prevented CBC versus an unnecessary CPM. The factor is derived from the threshold probability to develop a CBC at 10 years at which a patient would opt for CPM (e.g. 4.5%). The x-axis represents the threshold probability. Using a threshold probability of 4.5% implicitly means that CPM in 22 patients of whom one would develop a CBC if untreated is acceptable (21 unnecessary CPMs, harm to benefit ratio 1:21).

Table S10. Clinical utility of the 10-year contralateral breast cancer risk prediction model in non-*BRCA* tested patients. At the same probability threshold, the net benefit is exemplified in patients with family history (for avoiding unnecessary CPM) and patients without family history (performing necessary CPM).

		Family history		No family history	
Probability threshold p_t (%)	Unnecessary CPMs needed to prevent a CBC*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients
3.5	27.6	0.3	8.3	5.0	137.9
4.5	21.2	2.6	55.2	3.0	63.7
5.5	17.2	8.2	140.9	1.1	18.9

CPM: contralateral mastectomy; CBC: contralateral breast cancer;
*The number of unnecessary contralateral preventive mastectomies needed to prevent a CBC is calculated by: $(1-p_t)/p_t$

9. Formula to estimate the contralateral breast cancer risk in patients not tested for BRCA

The formula for the alternative model is reported below. Baseline survival estimates according to the model and time are:

$$S_0(5) = 0.982$$
$$S_0(10) = 0.965$$

And
Linear Predictor (LP) =
+ 0.108 – 0.002 × Age – 0.018 × Age' + 0.473 × I[Family history = Yes] – 0.143 × I[Nodal status = positive] – 0.041 × I[Size of PBC = (2,5] *cm*] + 0.108 × I[Size of PBC = greater than 5 *cm*] + 0.181 × I[Morphology of PBC = lobular including mixed] – 0.032 × I[Grade of PBC = moderately differentiated] – 0.135 × I[Grade of PBC = poorly/undifferentiated]– 0.248 × I[Chemotherapy = yes] – 0.034 × I[Radiotherapy to the breast = yes] + 0.522 × I[ER-negative without endocrine therapy] + 0.592 × I[ER-positive without endocrine] + 0.232 × I[HER2-negative without trastuzumab] + 0.082 × I[HER2-positive without trastuzumab]

Where Age' = max(Age – 50, 0)

10. Assessment of limited information of contralateral preventive mastectomy (CPM)

Information about CPM was not available in most studies. This lack of information may underestimate the cumulative CBC incidence because patients underwent CPM should not be considered to be at risk to develop CBC, though a small proportion of 1.3% of CBC was observed after CPM among *BRCA1* or *BRCA2*-related breast cancer patients[20]. We

investigated the impact of CPM on CBC cumulative incidence estimation in the BOSOM and EMC datasets, in which this information was complete and CPM was not within 3 months after first BC diagnosis, i.e. for 3,760 out of 3,793 and 3,390 out of 3,398, respectively. In these two studies, we compared the estimated cumulative incidence curves in which we applied censoring for CPM or considering in the risk set patients experiencing CPM at first primary BC or during the follow-up.

Figure S11 shows the cumulative incidence estimation of the two scenarios. As expected, the cumulative incidence was underestimated when we ignored the occurrence of CPM. However, there was only a small difference between the two curves: the estimated cumulative incidence at 10 years was 5.6% (95% CI: 4.9 – 6.4%) considering CPM, and 5.3% (4.6 – 6.0%) not considering CPM in the BOSOM dataset; and 5.7% (5.0 – 6.6%) considering CPM, and 5.6% (4.8 – 6.4%) not considering CPM in the EMC dataset. Therefore, although the CPM was not available for most studies, we concluded that the cumulative incidence of CBC was only slightly underestimated due to missing CPM information.

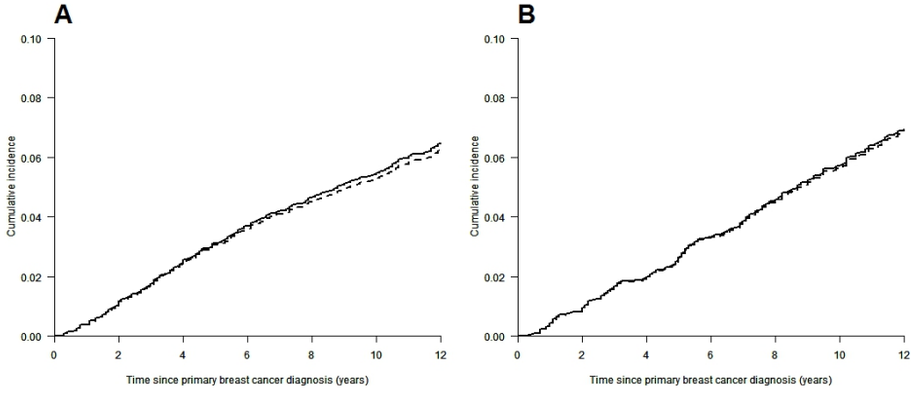
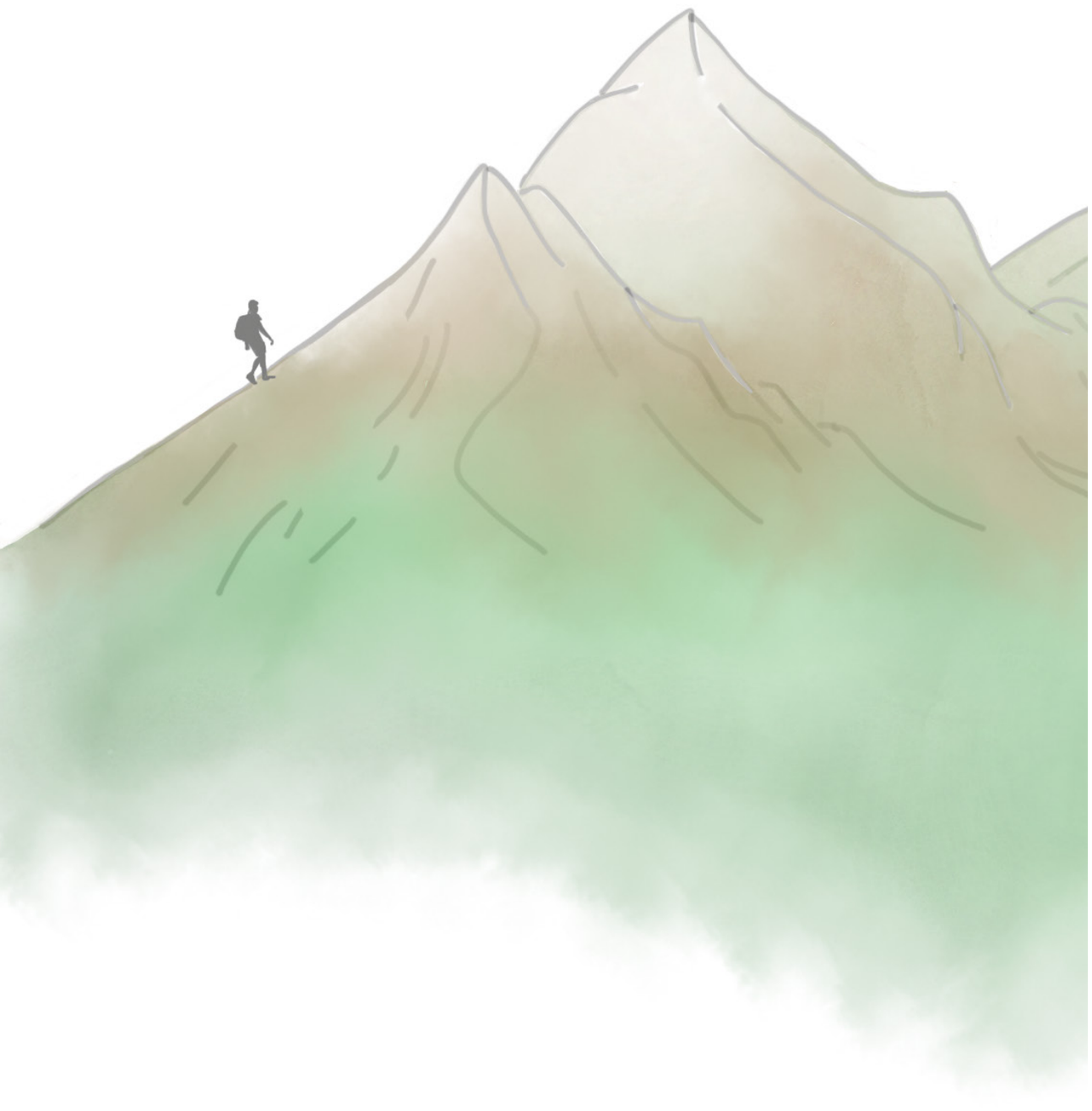


Figure S11: Assessment of inclusion of information of contralateral preventive mastectomy (CPM)**.** Panel A shows the contralateral breast cancer cumulative incidence curve in the BOSOM dataset. Panel B shows the contralateral breast cancer cumulative incidence curve in the EMC dataset. The bolded black lines indicate the estimated cumulative incidence curve censoring patients with CPM at first primary breast cancer or during the follow-up. The dotted lines indicate the estimated cumulative incidence curve considering patients with CPM still at risk during the follow-up.

REFERENCES

1. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A *et al*: **Association analysis identifies 65 new breast cancer risk loci**. *Nature* 2017, **551**(7678):92-94.
2. Schmidt MK, Tollenaar RA, de Kemp SR, Broeks A, Cornelisse CJ, Smit VT, Peterse JL, van Leeuwen FE, Van't Veer LJ: **Breast cancer survival and tumor characteristics in premenopausal women carrying the CHEK2*1100delC germline mutation**. *J Clin Oncol* 2007, **25**(1):64-69.
3. Schmidt MK, van den Broek AJ, Tollenaar RA, Smit VT, Westenend PJ, Brinkhuis M, Oosterhuis WJ, Wesseling J, Janssen-Heijnen ML, Jobsen JJ *et al*: **Breast Cancer Survival of BRCA1/BRCA2 Mutation Carriers in a Hospital-Based Cohort of Young Women**. *J Natl Cancer Inst* 2017, **109**(8).
4. Font-Gonzalez A, Liu L, Voogd AC, Schmidt MK, Roukema JA, Coebergh JW, de Vries E, Soerjomataram I: **Inferior survival for young patients with contralateral compared to unilateral breast cancer: a nationwide population-based study in the Netherlands**. *Breast Cancer Res Treat* 2013, **139**(3):811-819.
5. Kramer I, Schaapveld M, Oldenburg HSA, Sonke GS, McCool D, Van Leeuwen FE, van de Vijver KK, Russell NS, Linn SC, Siesling S *et al*: **The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype**. *J Natl Cancer Inst* In press.
6. Bouchardy C, Usel M, Verkooijen HM, Fioretta G, Benhamou S, Neyroud-Caspar I, Schaffar R, Vlastos G, Wespi Y, Schafer P *et al*: **Changing pattern of age-specific breast cancer incidence in the Swiss canton of Geneva**. *Breast Cancer Res Treat* 2010, **120**(2):519-523.
7. Riley RD, Lambert PC, Abo-Zaid G: **Meta-analysis of individual participant data: rationale, conduct, and reporting**. *BMJ* 2010, **340**:c221.
8. Resche-Rigon M, White IR, Bartlett JW, Peters SA, Thompson SG, Group P-IS: **Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data**. *Stat Med* 2013, **32**(28):4890-4905.
9. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG: **Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE**. *Stat Med* 2015, **34**(11):1841-1863.
10. White IR, Royston P: **Imputing missing covariate values for the Cox model**. *Stat Med* 2009, **28**(15):1982-1998.
11. Collins GS, Ogundimu EO, Altman DG: **Sample size considerations for the external validation of a multivariable prognostic model: a resampling study**. *Stat Med* 2016, **35**(2):214-226.
12. Blanche P, Dartigues JF, Jacqmin-Gadda H: **Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks**. *Stat Med* 2013, **32**(30):5381-5397.
13. Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, Riley RD: **Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model**. *J Clin Epidemiol* 2016, **69**:40-50.
14. Vickers AJ, Elkin EB: **Decision curve analysis: a novel method for evaluating prediction models**. *Med Decis Making* 2006, **26**(6):565-574.
15. Peirce CS: **The numerical measure of the success of predictions**. *Science* 1884, **4**(93):453-454.

16. Localio AR, Goodman S: **Beyond the usual prediction accuracy metrics: reporting results for clinical decision making**. *Ann Intern Med* 2012, **157**(4):294-295.
17. Vickers AJ, Cronin AM, Elkin EB, Gonen M: **Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers**. *BMC Med Inform Decis Mak* 2008, **8**:53.
18. Kerr KF, Brown MD, Zhu K, Janes H: **Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use**. *J Clin Oncol* 2016, **34**(21):2534-2540.
19. Van Calster B, Vickers AJ: **Calibration of risk prediction models: impact on decision-analytic performance**. *Med Decis Making* 2015, **35**(2):162-169.
20. van Sprundel TC, Schmidt MK, Rookus MA, Brohet R, van Asperen CJ, Rutgers EJ, Van't Veer LJ, Tollenaar RA: **Risk reduction of contralateral breast cancer and survival after contralateral prophylactic mastectomy in BRCA1 or BRCA2 mutation carriers**. *Br J Cancer* 2005, **93**(3):287-292.



Chapter 3

Prediction of contralateral breast cancer: External validation of risk calculators in 20 international cohorts

Breast cancer research and treatment. 2020 Jun;181(2):423-34[#].
https://link.springer.com/article/10.1007/s10549-020-05611-8

Daniele Giardiello
Michael Hauptmann
Ewout W. Steyerberg
Muriel A. Adank, Delal Akdeniz, Jannet C. Blom, Carl Blomqvist, Stig E. Bojesen, Manjeet K. Bolla, Mariël Brinkhuis, Jenny Chang-Claude, Kamila Czene, Peter Devilee, Alison M. Dunning, Douglas F. Easton, Diana M. Eccles, Peter A. Fasching, Jonine Figueroa, Henrik Flyger, Montserrat García-Closas, Lothar Haeberle, Christopher A. Haiman, Per Hall, Ute Hamann, John L. Hopper, Agnes Jager, Anna Jakubowska, Audrey Jung, Renske Keeman, Linetta B. Koppert, Iris Kramer, Diether Lambrechts, Loic Le Marchand, Annika Lindblom, Jan Lubiński, Mehdi Manoochehri, Luigi Mariani, Heli Nevanlinna, Hester S.A. Oldenburg, Saskia Pelders, Paul D.P. Pharoah, Mitul Shah, Sabine Siesling, Vincent T.H.B.M. Smit, Melissa C. Southey, William J. Tapper, Rob A.E.M. Tollenaar, Alexandra J. van den Broek, Carolien H.M. van Deurzen, Flora E. van Leeuwen, Chantal van Ongeval, Laura J. Van't Veer, Qin Wang, Camilla Wendt, Pieter J. Westenend, Maartje J. Hooning
Marjanka K. Schmidt

[#]A full list of authors and their affiliations appears on the journal's website

ABSTRACT

Background

Three tools are currently available to predict the risk of contralateral breast cancer (CBC). We aimed to compare the performance of the Manchester formula, CBCrisk, and PredictCBC in patients with invasive breast cancer (BC).

Methods

We analyzed data of 132,756 patients (4,682 CBC) from 20 international studies with a median follow-up of 8.8 years. Prediction performance included discrimination, quantified as a time-dependent Area-Under-the-Curve (AUC) at 5 and 10 years after diagnosis of primary BC, and calibration, quantified as the expected-observed (E/O) ratio at 5 and 10 years and the calibration slope.

Results

The AUC at 10 years was: 0.58 (95% confidence intervals [CI]: 0.57–0.59) for CBCrisk; 0.60 (95%CI: 0.59–0.61) for the Manchester formula; 0.63 (95%CI: 0.59–0.66) and 0.59 (95%CI: 0.56–0.62) for PredictCBC-1A (for settings where *BRCA1/2* mutation status is available) and PredictCBC-1B (for the general population), respectively. The E/O at 10 years: 0.82 (95%CI: 0.51–1.32) for CBCrisk; 1.53 (95%CI: 0.63–3.73) for the Manchester formula; 1.28 (95%CI: 0.63–2.58) for PredictCBC-1A and 1.35 (95%CI: 0.65–2.77) for PredictCBC-1B. The calibration slope was 1.26 (95%CI: 1.01–1.50) for CBCrisk; 0.90 (95%CI: 0.79–1.02) for PredictCBC-1A; 0.81 (95%CI: 0.63–0.99) for PredictCBC-1B, and 0.39 (95%CI: 0.34–0.43) for the Manchester formula.

Conclusions

Current CBC risk prediction tools provide only moderate discrimination and the Manchester formula was poorly calibrated. Better predictors and re-calibration are needed to improve CBC prediction and to identify low and high CBC risk patients for clinical decision making.

Keywords

Contralateral breast cancer, risk prediction, validation, clinical decision making

INTRODUCTION

A rising number of women with breast cancer (BC) are at risk to develop a new primary tumor in the contralateral breast (CBC) with consequently another cancer treatment and potentially less favorable prognosis[1]. Although CBC incidence is low (~0.4% per year) in the general BC population, contralateral preventive mastectomy (CPM) is increasing, also among women with low CBC risk[2-5].

Three tools are tools currently available to predict the risk of CBC, although probably none are widely used: 1) the Manchester formula; 2) CBCrisk, and 3) PredictCBC[6-8]. The Manchester group in the United Kingdom (UK) proposed a set of guidelines for counseling women about CPM[8]. Based on a systematic review of the literature, they devised a formula to estimate lifetime CBC risk based on age at first primary BC, family history of BC, estrogen-receptor (ER) status, diagnosis of ductal carcinoma in situ (DCIS), and oophorectomy.

The second tool, CBCrisk, was developed using data on 1,921 CBC cases and 5,763 matched controls with primary BC[7]. The model uses data on age at first BC diagnosis, age at first birth, first degree family history of BC, high-risk pre-neoplasia, breast density (obtained using the BI-RADS system), ER status, first BC type (pure invasive, pure DCIS, a mix of the two, unknown), and adjuvant endocrine therapy. External validation was performed using two independent studies in the United States (US) of 5,185 and 6,035 patients with 111 and 117 CBC events[7,9]. A web-based application provides individualized prediction of CBC risk[10].

Third, PredictCBC was developed, cross-validated and evaluated using data from 132,756 patients with first BC and 4,672 CBC events, as part of an international collaboration[5]. PredictCBC predicts CBC risk as a function of family history (first degree) of primary BC, and information of primary BC diagnosis: age, nodal status, size, grade, morphology, ER status, human epidermal growth factor receptor 2 (HER2) status, administration of adjuvant or neoadjuvant chemotherapy, adjuvant endocrine therapy, adjuvant trastuzumab therapy, and radiotherapy. Two versions were developed: PredictCBC version 1A includes presence or absence of a mutation in the *BRCA1* or *BRCA2* genes, an important determinant of CBC[5,11,12], while PredictCBC version 1B was developed for untested patients.

External validation in different studies is relevant to assess the prediction performance of prediction models[13]. Our aim was to perform a head-to-head comparison between CBCrisk, PredictCBC and the Manchester formula. We hereto used several large population- and hospital-based studies used to develop and cross-validate the PredictCBC models.

MATERIAL AND METHODS

External validation of CBCrisk and the Manchester formula was performed in 20 studies: four with individual patient data from the Netherlands (the Amsterdam Breast Cancer Study (ABCS), the Breast Cancer Outcome Study of Mutation carriers (BOSOM), the Erasmus MC Breast Cancer Registry (EMC), the Netherlands Cancer Registry (NCR)); and 16 other studies of the Breast Cancer Association Consortium (BCAC). The latter is an international consortium of 102 studies comprising 182,898 patients (data version: January 2017) with a primary BC diagnosed between 1939 and 2016[14]. Of these, 16 non-familial BC BCAC studies including invasive non-metastatic European-descent female patients with first primary invasive BC diagnosed from 1990 onwards, and with at least 10 CBC events, were included in the analyses[14]. Details about studies and patient selection, and data imputation were described previously[5].

The outcome was in situ or invasive metachronous CBC. Follow-up started 3 months after invasive first primary BC diagnosis, to exclude synchronous CBCs, and ended at date of CBC, distant metastasis (but not at loco-regional relapse), CPM or last date of follow-up (due to death, being lost to follow-up, or end of study), whichever occurred first. In the BCAC, 27,155 patients were recruited more than 3 months after diagnosis of the first primary BC (prevalent cases); for these patients, follow-up started at date of recruitment (left truncation). Distant metastasis and death due to any cause were competing events.

The Manchester formula provides an estimate of a woman's individual life-time CBC risk. To assess the prediction performance, we translated the life-time CBC risk to 5- and 10-year CBC risks (see **Supplementary Material**). The predictors included in the CBC risk estimation in the Manchester formula, CBCrisk and PredictCBC models are provided in **Table 1**. Predictors that were sporadically missing were multiply imputed as described elsewhere[5].

Statistical analysis

Discrimination, the ability of the model to differentiate between patients who experienced CBC and those who did not, was calculated by time-dependent Area-Under-the-Curve (AUCs) based on Inverse Censoring Probability Weighting at 5 and 10 years[15,16]. Values of AUCs close to 1 indicate good discrimination while values close to 0.5 indicate poor discrimination (a coin flip). Calibration is the agreement between observed and predicted risk and is commonly characterized by calibration-in-the-large and slope statistic. Calibration-in-the-large characterizes the overall difference between the observed and predicted risks. It was calculated using the expected/observed (E/O) ratio. An E/O less than 1 indicates that the model systematically underestimates CBC risk, while an E/O above 1 indicates that the model systematically overestimates CBC risk. The expected number of cases was calculated by summing the individual predicted probabilities at 5 and 10 years, based on the patient-specific covariate values[17]. The observed number of cases was estimated by the non-parametric CBC cumulative incidence at 5 and 10 years. The calibration slope was estimated using a Fine and Gray regression model using the linear predictor of the prediction tools. The linear predictor was vs constructed as the sum of the factors included in each model weighted by the corresponding regression coefficients (or parameters), and then computed in the validation dataset exactly as reported for the development set. The calibration slope is determined as the regression coefficient for this linear predictor when fitted as a single covariate in a regression model of disease outcome in the validation dataset. A well-calibrated model should have a calibration slope of 1; slopes < 1 indicate that coefficients were too optimistic for the validation setting[18]. Calibration results were graphically displayed.

Table 1: Predictors included in current contralateral breast cancer risk prediction tools

List of predictors	CBCrisk§	Manchester formula†	PredictCBC version 1A‡	PredictCBC version 1B‡
Age at diagnosis	✓	✓	✓	✓
Age at first birth	✓			
First-degree family history	✓	✓	✓	✓
BRCA1/2 germline mutation		✓	✓	
First breast cancer behavior type*	✓	✓		
Lymph node status			✓	✓
Breast density	✓			
Tumor size			✓	✓
Morphology			✓	✓
Tumor grade			✓	✓
High risk pre-neoplasia	✓			
ER status	✓	✓	✓	✓
HER2 status			✓	✓
Chemotherapy			✓	✓
Endocrine therapy	✓		✓	✓
Radiation to the breast			✓	✓
Trastuzumab			✓	✓
Oophorectomy under 40 years		✓		

* Contralateral breast cancer risk was calculated including women diagnosed with ductal carcinoma in situ;
§Chowdhury M, Euhus D, Onega T, Biswas S, Choudhary PK (2017) A model for individualized risk Abbreviation: ER: estrogen receptor status; HER2: human epidermal growth factor receptor 2.
prediction of contralateral breast cancer. Breast Cancer Res Treat 161 (1):153-160.
†Basu NN, Ross GL, Evans DG, Barr L (2015) The Manchester guidelines for contralateral risk-reducing mastectomy. World J Surg Oncol 13:237
‡ Giardiello D, Steyerberg EW, Hauptmann M, Adank MA, Akdeniz D, Blomqvist C, Bojesen SE, Bolla MK, Brinkhuis M, Chang-Claude J, Czene K, Devilee P, Dunning AM, Easton DF, Eccles DM, Fasching PA, Figueroa J, Flyger H, Garcia-Closas M, Haeberle L, Haiman CA, Hall P, Hamann U, Hopper JL, Jager A, Jakubowska A, Jung A, Keeman R, Kramer I, Lambrechts D, Le Marchand L, Lindblom A, Lubinski J, Manoochehri M, Mariani L, Nevanlinna H, Oldenburg HSA, Pelders S, Pharoah PDP, Shah M, Siesling S, Smit V, Southey MC, Tapper WJ, Tollenaar R, van den Broek AJ, van Deurzen CHM, van Leeuwen FE, van Ongeval C, Van't Veer LJ, Wang Q, Wendt C, Westenend PJ, Hooning MJ, Schmidt MK (2019) Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res 21 (1):144. doi:10.1186/s13058-019-1221-1

Analyses were stratified by geographic groups of studies, since stratification by individual studies would provide too few events in some strata[13,19,5]. To allow for heterogeneity across multiple studies, random-effect meta-analyses were performed. We calculated 95% confidence intervals (CI) and 95% prediction intervals (PI), which indicate the likely range for prediction accuracy of the model in a new dataset, for discrimination and calibration measures. A sensitivity analysis was performed to check the consistency of CBCrisk performance measures when metachronous CBC was defined as an event after 6 instead of 3 months since the first BC diagnosis. More details are provided in the **Supplementary Material**. All analyses were implemented using SAS (SAS Institute Inc., NC, USA) and R software[20].

RESULTS

We included 132,756 patients from 20 studies who experienced 4,862 CBC events during a median follow-up of 8.8 years. The main patient and clinical characteristics across studies and geographic areas are shown in **Table 2**.

The AUCs at 5 and 10 years was around 0.6: 0.59 (95% CI: 0.57–0.61; 95% PI: 0.54–0.64) and 0.58 (95% CI: 0.57–0.59; 95% PI: 0.55–0.61) for CBCrisk (**Figure 1**); 0.61 (95% CI: 0.60–0.62; 95% PI: 0.59–0.63) and 0.60 (95% CI: 0.59–0.61; 95% PI: 0.58–0.62) for the Manchester formula (**Figure 2**). The E/O ratio at 5 and 10 years was close to 1 for all models: 0.86 (95% CI: 0.50–1.46; 95% PI: 0.20–3.75) and 0.82 (95% CI: 0.51–1.32; 95% PI: 0.21–3.14) for CBCrisk (**Table 3**); 1.54 (95% CI: 0.61–3.92; 95% PI: 0.11–20.72, **Table 4**), and 1.53 (95% CI: 0.63–3.73; 95% PI: 0.13–18.52) for the Manchester formula (**Table 4**); 1.26 (95% CI: 0.57–2.77; 95% PI: 0.14–11.34), and 1.28 (95% CI: 0.63–2.58; 95% PI: 0.18–9.18) for PredictCBC-1A (**Table 5**); 1.33 (95% CI: 0.59–2.99, 95% PI: 0.14–12.76), 1.35 (95% CI: 0.65–2.77; 95% PI: 0.19–10.24) for PredictCBC-1B (**Table 5**)[5]. The calibration slope was close to 1 for CBCrisk (1.26, 95% CI: 1.01–1.50 and 95% PI: 1.01–1.50, **Table 3-5**), and PredictCBC-1A and 1B 0.90 (95% CI: 0.79–1.02; 95% PI: 0.73–1.08), and 0.81 (95% CI: 0.63–0.99; 95% PI: 0.50–1.12) (**Table 5**), while prognostic effects were far too large for the Manchester formula (slope: 0.39, 95% CI: 0.34–0.43, 95% PI: 0.34–0.43, **Table 4-5**). Calibration plots of CBCrisk at 5 and 10 years are shown in **Supplementary Figure 1** and **Supplementary Figure 2**. As reported previously[5], the AUCs at 5 and 10 years for PredictCBC-1A were 0.63 (95% CI: 0.58–0.67, 95% PI: 0.52–0.74), and 0.63 (95% CI: 0.59–0.66, 95% PI: 0.53–0.72), respectively; for PredictCBC-1B 0.59 (CI: 0.54–0.63, 95% PI: 0.46–0.71, **Table 5**), and 0.59 (95% CI: 0.56–0.62, 95% PI: 0.52–0.66, **Table 5**), respectively.

Table 2: Description of main patient and clinical factors used for evaluation of the models and formula ‡

Study* / Geographic area	Europe - other§	Europe - Scandinavia	Europe - United Kingdom	Netherlands - BOSOM	Netherlands - EMC	Netherlands - NCR	United States and Australia
N	15,183	12,928	11,921	3,760	3,390	83,138	2,436
Age at first diagnosis, *years* (%)							
<30	152 (1.0)	46 (0.4)	156 (1.3)	108 (2.9)	46 (1.4)	388 (0.5)	41 (1.7)
30-39	1,252 (8.2)	489 (3.8)	1,811 (15.2)	842 (22.4)	374 (11.0)	4,241 (5.1)	494 (20.3)
40+	13,779 (90.8)	12,393 (95.9)	9,954 (83.5)	2,810 (74.7)	2,970 (87.6)	78,509 (94.4)	1,901 (78.0)
Age at first birth = unknown (%)	15,183 (100.0)	12,928 (100.0)	11,921 (100.0)	3,760 (100.0)	3,390 (100.0)	83138 (100.0)	2,436 (100.0)
Family history (%)							
Yes	2,123 (14.0)	818 (6.3)	1,371 (11.5)	737 (19.6)	591 (17.4)	0 (0.0)	319 (13.1)
No	8,057 (53.1)	3,158 (24.4)	8,210 (68.9)	1,177 (31.3)	2,482 (73.2)	0 (0.0)	1,498 (61.5)
Unknown	5,003 (33.0)	8,952 (69.2)	2,340 (19.6)	1,846 (49.1)	317 (9.4)	83,138 (100.0)	619 (25.4)
First BC type = Pure invasive (%)	15,183 (100.0)	12,928 (100.0)	11,921 (100.0)	3,760 (100.0)	3,390 (100.0)	83,138 (100.0)	2,436 (100.0)
Breast density = unknown (%)	15,183 (100.0)	12,928 (100.0)	11,921 (100.0)	3,760 (100.0)	3,390 (100.0)	83,138 (100.0)	2,436 (100.0)
ER status (%)							
Negative	3,387 (22.3)	1,746 (13.5)	1,718 (14.4)	896 (23.8)	842 (24.8)	14,591 (17.6)	445 (18.3)
Positive	10,071 (66.3)	9,401 (72.7)	7,175 (60.2)	2,024 (53.8)	2,427 (71.6)	64,790 (77.9)	1,572 (64.5)
Unknown	1,725 (11.4)	1,781 (13.8)	3,028 (25.4)	840 (22.3)	121 (3.6)	3,757 (4.5)	419 (17.2)
High risk pre-neoplasia = unknown (%)	15,183 (100.0)	12,928 (100.0)	11,921 (100.0)	3,760 (100.0)	3,390 (100.0)	83,138 (100.0)	2,436 (100.0)
Anti-estrogen therapy (%)							
Yes	7,868 (51.8)	6,434 (49.8)	8,712 (73.1)	809 (21.5)	1,559 (46.0)	40,214 (48.4)	363 (14.9)
No	4,570 (30.1)	1,947 (15.1)	2,046 (17.2)	2,739 (72.8)	1,821 (53.7)	42,924 (51.6)	8 (0.3)
Unknown	2,745 (18.1)	4,547 (35.2)	1,163 (9.8)	212 (5.6)	10 (0.3)	0 (0.0)	2,065 (84.8)
CBC cumulative incidence (%)							
3-year (95% CI)	1.0 (0.8 - 1.2)	0.7 (0.5 - 0.9)	0.5 (0.3 - 0.7)	1.7 (1.3 - 2.1)	1.7 (1.2 - 2.1)	1.3 (1.2 - 1.4)	1.8 (0.8 - 2.8)
5-year (95% CI)	1.6 (1.4 - 1.9)	1.0 (0.8 - 1.3)	1.0 (0.8 - 1.3)	3.0 (2.5 - 3.6)	2.6 (2.1 - 3.2)	2.4 (2.3 - 2.5)	2.8 (1.7 - 3.8)
10-year (95% CI)	3.5 (3.1 - 3.9)	2.1 (1.7 - 2.4)	1.3 (1.0 - 1.5)	5.5 (4.7 - 6.2)	5.7 (4.9 - 6.6)	4.6 (4.5 - 4.8)	4.1 (3.0 - 5.3)

‡More details about the main patient and clinical characteristics by study are available in the supplementary information of [5]
Abbreviations:
* The studies denoted with Europe and United States and Australia are part of the Breast Cancer Association Consortium
§ Europe - other geographic area included studies from Belgium (1), Germany (2), Netherlands (2) and Poland (2).
BOSOM: Breast Cancer Outcome Study of Mutation carriers; EMC: Erasmus Medical Center; NCR: Netherlands Cancer Registry
BC: breast cancer; ER: estrogen receptor; CBC: contralateral breast cancer; CI: confidence interval;

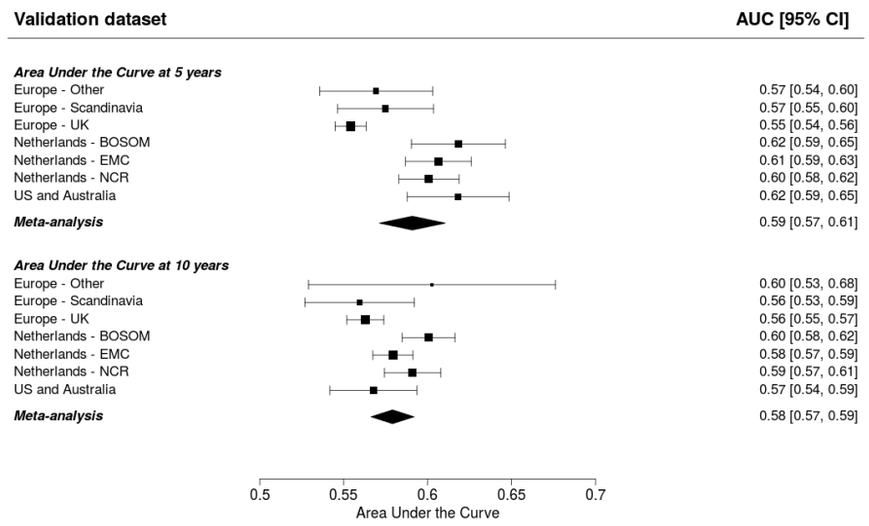


Figure 1: Prediction performance of the CBCrisk model (Chowdhury et al. [7]). The upper and lower panel show the discrimination assessed by a time-dependent Area-Under-the-Curve at 5 and 10 years, respectively. The black squares indicate the estimated accuracy of a model built on all remaining studies or geographic areas. The black horizontal lines indicate the corresponding 95% confidence intervals of the estimated accuracy (interval whiskers). The black diamonds indicate the mean with the corresponding 95% confidence interval of the predictive accuracy.

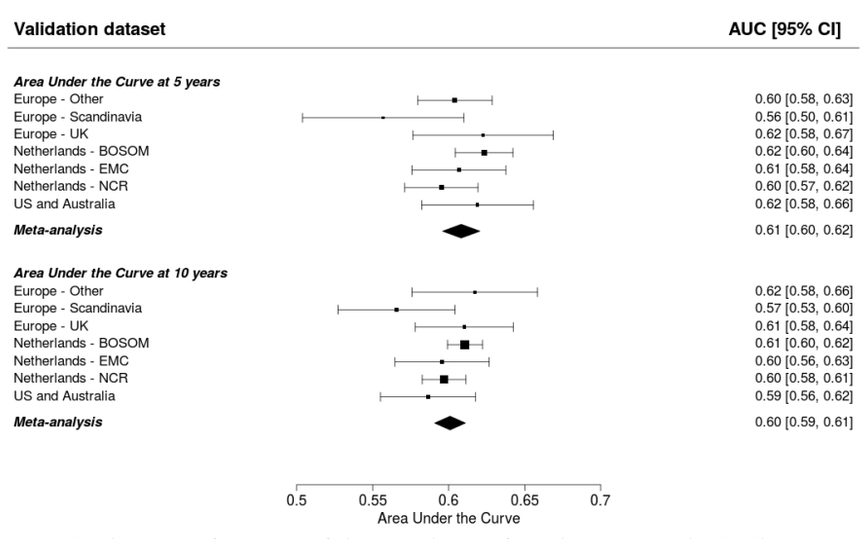


Figure 2: Prediction performance of the Manchester formula (Basu et al. [8]) The upper and lowerpanel show the discrimination assessed by a time-dependent Area-Under-the-Curve at 5 and 10 years, respectively. The black squares for each dataset indicate the estimated accuracy of a model built on all remaining studies or geographic areas. The black horizontal lines indicate the corresponding 95% confidence intervals of the estimated accuracy (intervalwhiskers). The black diamonds indicate the mean with the corresponding 95% confidence interval of the predictive accuracy.

Table 3: Calibration performance of the CBCrisk model[§]

Validation dataset	E/O ratio at 5 years (95% CI)	E/O ratio at 10 years (95% CI)	Calibration slope (95% CI)
Europe - Other	0.87 (076 - 0.98)	0.75 (0.68 - 0.81)	1.11 (0.40 - 1.83)
Europe - Scandinavia	1.59 (1.28 - 1.91)	1.23 (1.08 - 1.38)	0.86 (0.16 - 1.57)
Europe - UK	1.35 (1.38 - 2.17)	1.82 (1.53 - 2.11)	0.85 (-0.03 - 1.73)
Netherlands - BOSOM	0.45 (0.37 - 0.53)	0.50 (0.43 - 0.57)	1.34 (0.76 - 1.93)
Netherlands - EMC	0.48 (0.38 - 0.57)	0.43 (0.37 - 0.50)	1.19 (0.65 - 1.73)
Netherlands - NCR	0.57 (0.54 - 0.59)	0.54 (0.52 - 0.56)	1.40 (1.11 - 1.68)
US and Australia	0.43 (0.33 - 0.54)	0.56 (0.45 - 0.67)	1.13 (0.25 - 2.00)
Meta-analysis	0.86 (0.50 - 1.46)	0.82 (0.51 - 1.32)	1.26 (1.01 – 1.50)
95% PI	0.20 - 3.75	0.21 - 3.14	1.01 - 1.50

Abbreviations: E/O: expected-observed; CI: confidence interval; UK: United Kingdom; BOSOM: Breast Cancer Outcome Study of Mutation carriers; EMC: Erasmus Medical; Center NCR: Netherlands Cancer Registry; PI: prediction interval

[§]Chowdhury M, Euhus D, Onega T, Biswas S, Choudhary PK (2017) A model for individualized risk prediction of contralateral breast cancer. Breast Cancer Res Treat 161 (1):153-160.

Table 4: Calibration performance of the Manchester formula[§]

Validation dataset	E/O ratio at 5 years (95% CI)	E/O ratio at 10 years (95% CI)	Calibration slope (95% CI)
Europe - Other	1.64 (1.44 - 1.85)	1.46 (1.34 - 1.58)	0.40 (0.29 - 0.50)
Europe - Scandinavia	2.61 (2.09 - 3.12)	2.11 (1.85 - 2.37)	0.35 (0.13 - 0.57)
Europe - UK	3.34 (2.60 - 4.08)	3.49 (2.93 - 4.05)	0.42 (0.23 - 0.61)
Netherlands - BOSOM	0.81 (0.66 - 0.96)	0.92 (0.79 - 1.05)	0.45 (0.33 - 0.56)
Netherlands - EMC	0.94 (0.75 - 1.14)	0.87 (0.75 - 1.00)	0.35 (0.21 - 0.49)
Netherlands - NCR	1.00 (0.95 - 1.04)	1.01 (0.98 - 1.05)	0.37 (0.33 - 0.42)
US and Australia	0.77 (0.58 - 0.96)	1.02 (0.82 - 1.23)	0.51 (0.33 - 0.68)
Meta-analysis	1.54 (0.61 - 3.92)	1.53 (0.63 - 3.73)	0.39 (0.34 - 0.43)
95% PI	0.11 - 20.72	0.13 - 18.52	0.34 - 0.43

Abbreviations: E/O: expected-observed; CI: confidence interval; UK: United Kingdom; BOSOM: Breast Cancer Outcome Study of Mutation carriers; EMC: Erasmus Medical; Center NCR: Netherlands Cancer Registry; PI: prediction interval

[§]Basu NN, Ross GL, Evans DG, Barr L (2015) The Manchester guidelines for contralateral risk-reducing mastectomy. World J Surg Oncol 13:237

Sensitivity analysis showed that the performance measures of CBCrisk did not change when metachronous CBC was defined after 6 months since first BC diagnosis (see **Supplementary Materials, Supplementary Table 1-2 and Supplementary Figure 3**).

Table 5: Summary of prediction performance of CBCrisk, Manchester formula, and PredictCBC version 1A and version 1B with the corresponding 95% prediction intervals (PI).

Characteristics	CBCrisk[§]	Manchester formula[†]	PredictCBC version 1A[‡*]	PredictCBC version 1B[‡*]
Discrimination				
AUC at 5 years (95% PI)	0.59 (0.54 - 0.64)	0.61 (0.59 - 0.63)	0.63 (0.52 - 0.74)	0.59 (0.46 - 0.71)
AUC at 10 years (95% PI)	0.58 (0.55 - 0.61)	0.60 (0.58 - 0.62)	0.63 (0.53 - 0.72)	0.59 (0.52 - 0.66)
Calibration				
E/O ratio at 5 years (95% PI)	0.86 (0.20 - 3.75)	1.54 (0.11 - 20.72)	1.26 (0.14 - 11.34)	1.33 (0.14 - 12.76)
E/O ratio at 10 years (95% PI)	0.82 (0.21 - 3.14)	1.53 (0.13 - 18.52)	1.28 (0.18 - 9.18)	1.35 (0.19 - 10.24)
Slope (95% PI)	1.26 (1.01 - 1.50)	0.39 (0.34 - 0.43)	0.90 (0.73 - 1.08)	0.81 (0.50 - 1.12)

Abbreviations: AUC: Area under the curve; PI: prediction interval
[§]Chowdhury M, Euhus D, Onega T, Biswas S, Choudhary PK (2017) A model for individualized risk prediction of contralateral breast cancer. Breast Cancer Res Treat 161 (1):153-160.
[†]Basu NN, Ross GL, Evans DG, Barr L (2015) The Manchester guidelines for contralateral risk-reducing mastectomy. World J Surg Oncol 13:237
[‡] Giardiello D, Steyerberg E, Hauptmann M, et al. (2019) Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res. doi:10.1186/s13058-019-1221-1, Figure 1 and Figure S5
[*]version 1A includes *BRCA* mutation status as a variable while 1B does not.

DISCUSSION

Accurate CBC risk predictions are essential in clinical decision making around CPM or tailored surveillance among patients with first primary BC. In particular, overestimation of risk can lead to recommending CPM among BC patients with low risks. Underestimation can lead to suboptimal surveillance or hesitance about recommending CPM for patients with substantial risk. Using individual patient data from multiple studies with long follow-up, we externally evaluated the prediction performance accuracy of CBCrisk, a tool developed and validated to provide individualized CBC risk prediction, and the Manchester formula, a heuristically derived calculation of CBC lifetime risk[6,8,7,9]. In addition, the availability of different European-descendent studies allowed heterogeneity in the performance by geographic area to be assessed.

CBCrisk under-predicted the risk of CBC and had moderate discrimination ability with considerable heterogeneity between studies. The Manchester formula was empirically derived from a systematic review, and its discrimination accuracy was higher than CBCrisk. This may be explained by the inclusion of *BRCA1/2* mutation carrier information, an important determinant of CBC risk[21]. With the same large individual patient data sets, PredictCBC models had been developed and validated[5]. In particular, PredictCBC version 1A includes information of *BRCA1/2* mutation carriers and extensive information about the primary BC including treatments. The discrimination of all three prediction models was moderate, with AUC values around 0.6.

CBCrisk was previously externally validated using two independent clinical studies from Johns Hopkins University (JH) and MD Anderson Cancer Center (MDA) in the US[9]. Discrimination ability was 0.61 and 0.65 at 3 years, and 0.62 and 0.61 at 5 years for JH and MDA, respectively. The risk of CBC was overestimated in JH with E/O ratios of 2.02 and 1.56 at 3 and 5 years, while underestimated in MDA with E/O ratios of 0.61 and 0.62, respectively.

The considerable heterogeneity in all CBC risk calculators, especially in the CBCrisk and the Manchester formula, reflects the different CBC incidences in every study[13]. Another potential source of heterogeneity is the carrier frequency of germline mutations associated with CBC that may vary among studies, especially in the CBC calculators not including information of *BRCA1/2* mutation as CBCrisk and the PredictCBC-1B[22]. In addition, heterogeneity may be due to the different proportions of the use of (neo) adjuvant systemic therapies explained by the different distribution of tumor subtypes among studies[4]. Besides, inter-observer variation in pathological examination of BC among studies may lead to different adjuvant systemic therapy advice and, consequently, prediction of CBC risk[23]. Variation in prediction performance and limited generalizability of CBC risk calculators can also be partially explained by differences in how predictors are measured among studies[24,25]. For example, lack of family history knowledge may lead to uncertainty in risk prediction and varies according to demographics of the patients[26]. In particular, if in some studies BC patients misreported information about family history, the CBC risk would be over(under)estimated causing inappropriate decision-making regarding CPM or tailored surveillance. Some limitations of our study must be recognized. Firstly, our dataset, while large, had missing data for three covariates that were used in the CBCrisk model: breast density, age at first birth, and high risk preneoplasia. The authors of CBCrisk estimated the relative risks for patients with the unknown characteristics, but the use of the missing indicator variable is suboptimal compared to having the prognostic information available. It may lead to over or under-estimation of absolute CBC risk[27]. For this reason, we suggest that it is preferable to use multiple imputation of missing data, as is done in the PredictCBC models[28,29]. In addition, investigation of the potential source of model misspecification due to possible different definitions or measurement error was not possible[30-32].

In conclusion, current statistical risk prediction models and heuristic formulas provided moderate CBC individualized prediction performance. Careful re-calibration is required before considering these models for clinical decision making. A more direct comparison between the current CBC risk prediction models using a large external dataset with complete information on all factors included in all CBC prediction models would be ideal, but is currently unavailable. There is an ongoing debate about improvements of clinical prediction performance using machine learning approaches compared to

standard regression approaches for risk prediction[33,34]. However, irrespective of the methodology, better predictors are needed to predict CBC more accurately. Deeper biological insights and potential inclusion of other genetic markers such as *CHEK2* c.1100del mutation status and polygenic risk scores based on common genetic variants may improve CBC risk prediction, although rare mutations are unlikely to contribute substantially to CBC risk in the general population[35,36]. Life-style factors such as body mass index, alcohol consumption, and smoking also may help to better stratify high and low CBC risk patients even though these factors are difficult to measure accurately. Moreover, breast density may be important. More detailed information about adjuvant systemic therapies may better identify patients with low and high CBC risk since chemotherapy and especially endocrine therapy reduce CBC risk[4]. After extension and further external validation of prediction models for CBC risk, investigation of their potential clinical utility is an important future step.

Acknowledgements

We thank all individuals who took part in these studies and all researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. ABCFS thank Maggie Angelakos, Judi Maskiell, Gillian Dite. ABCS and BOSOM thanks all the collaborating hospitals and pathology departments and many individual that made this study possible, specifically, we wish to acknowledge: Annegien Broeks, Sten Cornelissen, Frans Hogervorst, Laura van 't Veer, Floor van Leeuwen, Emiel Rutgers. EMC thanks J.C. Blom-Leenheer, P.J. Bos,C.M.G. Crepin and M. van Vliet for data management. CGPS thanks staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. HEBCS thanks Taru A. Muranen, Kristiina Aittomäki, Karl von Smitten, Irja Erkkilä. KARMA thanks the Swedish Medical Research Counsel. LMBC thanks Gilian Peuteman, Thomas Van Brussel, EvyVanderheyden and Kathleen Corthouts. MARIE thanks Petra Seibold, Dieter Flesch-Janys, Judith Heinz, Nadia Obi, Alina Vrieling, Sabine Behrens, Ursula Eilber, Muhabbet Celik, Til Olchers and Stefan Nickels. ORIGO thanks E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The authors thank the registration team of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry as well as IKNL staff for scientific advice. PBCS thanks Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner. The ethical approval for the POSH study is MREC /00/6/69, UKCRN ID: 1137. We thank the SEARCH team.

Funding

This work is supported by the Alpe d'HuZes/Dutch Cancer Society (KWF Kankerbestrijding) project 6253.

BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report.

The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. BOSOM was supported by the Dutch Cancer Society grant numbers DCS-NKI 2001-2423, DCS-NKI 2007-3839, and DCSNKI 2009-4363; the Cancer Genomics Initiative; and notary office Spier & Hazenberg for the coding procedure. The EMC was supported by grants from Alpe d'HuZes/Dutch Cancer Society NKI2013-6253 and from Pink Ribbon 2012.WO39.C143. The HEBCS was financially supported by the Helsinki University Hospital Research Fund, the Finnish Cancer Society, and the Sigrid Juselius Foundation.

Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantzows foundation. The KARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. LMBC is supported by the 'Stichting tegen Kanker'. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419, 110826, 110828], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MEC was support by NIH grants CA63464, CA54281, CA098758, CA132839 and CA164973. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. Genotyping for PLCO was supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The POSH study is funded by Cancer Research UK (grants C1275/A11699, C1275/C22524,

C1275/A19187, C1275/A15956 and Breast Cancer Campaign 2010PR62, 2013PR044. PROCAS is funded from NIHR grant PGfAR 0707-10031. SEARCH is funded by Cancer Research UK [C490/A10124, C490/A16561] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. The University of Cambridge has received salary support for PDPP from the NHS in the East of England through the Clinical Academic Reserve. SKKDKFZS is supported by the DKFZ. The SZBCS (Szczecin Breast Cancer Study) was supported by Grant PBZ_KBN_122/P05/2004 and The National Centre for Research and Development (NCBR) within the framework of the international ERA-NET TRANSAN JTC 2012 application no. Cancer 12-054 (Contract No. ERA-NET-TRANSCAN / 07/2014).

Compliance with ethical standards

Funding: This work is supported by the Alpe d'HuZes/Dutch Cancer Society (KWF Kankerbestrijding) project 6253.

Conflict of interest: Author DG, MH, EW, MAA, DA, JCB, CB, SEB, MKB, JCC, KC, PD, AMD, DFE, JF, HF, MGC, LH, CAH, PH, UH, JLH, AJ, AJ2, AJ3, RK, LBK, IK, DL, LLN, AL, JL, MM, LM, HN, HSAO, SP, PDPP, MS, SS, VTHBMS, MCS, WJT, RAEMT, AJvdB, CHMvD, FEvL, CvO, LvV, QW, CW, PJW, MJH declares that he has no conflict of interest. Author DMM declares that she receives a lecture fee from Pierre Fabre and personal fees for consultancy from Astra Zeneca. Author PAF reports grants from Novartis, grants from Biontech, personal fees from Novartis, personal fees from Roche, personal fees from Pfizer, personal fees from Celgene, personal fees from Daiichi-Sankyo, personal fees from TEVA, personal fees from Astra Zeneca, personal fees from Merck Sharp & Dohme, personal fees from Myelo Therapeutics, personal fees from Macrogenics, personal fees from Eisai, personal fees from Puma, grants from Cepheid.

Ethical approval: all procedures performed in studies involving human participants were in accordance with the ethical standards of international, national, and institutional research committees and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: informed consent was obtained from all individual participants included in the study.

REFERENCES

1. Langballe R, Frederiksen K, Jensen MB, Andersson M, Cronin-Fenton D, Ejlertsen B, Mellemkjaer L (2018) Mortality after contralateral breast cancer in Denmark. Breast Cancer Res Treat 171 (2):489-499. doi:10.1007/s10549-018-4846-3
2. Xiong Z, Yang L, Deng G, Huang X, Li X, Xie X, Wang J, Shuang Z, Wang X (2018) Patterns of Occurrence and Outcomes of Contralateral Breast Cancer: Analysis of SEER Data. J Clin Med 7 (6). doi:10.3390/jcm7060133
3. Wong SM, Freedman RA, Sagara Y, Aydogan F, Barry WT, Golshan M (2017) Growing Use of Contralateral Prophylactic Mastectomy Despite no Improvement in Long-term Survival for Invasive Breast Cancer. Ann Surg 265 (3):581-589. doi:10.1097/SLA.0000000000001698
4. Kramer I, Schaapveld M, Oldenburg HSA, Sonke GS, McCool D, van Leeuwen FE, Van de Vijver KK, Russell NS, Linn SC, Siesling S, Menke-van der Houven van Oordt CW, Schmidt MK (2019) The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype. J Natl Cancer Inst. doi:10.1093/jnci/djz010
5. Giardiello D, Steyerberg EW, Hauptmann M, Adank MA, Akdeniz D, Blomqvist C, Bojesen SE, Bolla MK, Brinkhuis M, Chang-Claude J, Czene K, Devilee P, Dunning AM, Easton DF, Eccles DM, Fasching PA, Figueroa J, Flyger H, Garcia-Closas M, Haeberle L, Haiman CA, Hall P, Hamann U, Hopper JL, Jager A, Jakubowska A, Jung A, Keeman R, Kramer I, Lambrechts D, Le Marchand L, Lindblom A, Lubinski J, Manoochehri M, Mariani L, Nevanlinna H, Oldenburg HSA, Pelders S, Pharoah PDP, Shah M, Siesling S, Smit V, Southey MC, Tapper WJ, Tollenaar R, van den Broek AJ, van Deurzen CHM, van Leeuwen FE, van Ongeval C, Van't Veer LJ, Wang Q, Wendt C, Westenend PJ, Hooning MJ, Schmidt MK (2019) Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res 21 (1):144. doi:10.1186/s13058-019-1221-1
6. O'Donnell M (2018) Estimating Contralateral Breast Cancer Risk. Current Breast Cancer Reports 10 (2):91-97
7. Chowdhury M, Euhus D, Onega T, Biswas S, Choudhary PK (2017) A model for individualized risk prediction of contralateral breast cancer. Breast Cancer Res Treat 161 (1):153-160. doi:10.1007/s10549-016-4039-x
8. Basu NN, Ross GL, Evans DG, Barr L (2015) The Manchester guidelines for contralateral risk-reducing mastectomy. World J Surg Oncol 13:237. doi:10.1186/s12957-015-0638-y
9. Chowdhury M, Euhus D, Arun B, Umbricht C, Biswas S, Choudhary P (2018) Validation of a personalized risk prediction model for contralateral breast cancer. Breast Cancer Res Treat. doi:10.1007/s10549-018-4763-5
10. Chowdhury M, Euhus D, Onega T, Choudhary P (2017) CBCRisk: Contralateral Breast Cancer (CBC) Risk Predictor.
11. van den Broek AJ, van 't Veer LJ, Hooning MJ, Cornelissen S, Broeks A, Rutgers EJ, Smit VT, Cornelisse CJ, van Beek M, Janssen-Heijnen ML, Seynaeve C, Westenend PJ, Jobsen JJ, Siesling S, Tollenaar RA, van Leeuwen FE, Schmidt MK (2016) Impact of Age at Primary Breast Cancer on Contralateral Breast Cancer Risk in BRCA1/2 Mutation Carriers. J Clin Oncol 34 (5):409-418. doi:10.1200/JCO.2015.62.3942
12. Malone KE, Begg CB, Haile RW, Borg A, Concannon P, Tellhed L, Xue S, Teraoka S, Bernstein L, Capanu M, Reiner AS, Riedel ER, Thomas DC, Mellemkjaer L, Lynch CF, Boice JD, Jr., Anton-Culver H, Bernstein JL (2010) Population-based study of the risk of second primary contralateral breast cancer associated with carrying a mutation in BRCA1 or BRCA2. J Clin Oncol 28 (14):2404-2410. doi:10.1200/JCO.2009.24.2495

13. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW (2016) Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol 79:76-85. doi:10.1016/j.jclinepi.2016.05.007
14. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A, Bolla MK, Wang Q, Tyrer J, Dicks E, Lee A, Wang Z, Allen J, Keeman R, Eilber U, French JD, Qing Chen X, Fachal L, McCue K, McCart Reed AE, Ghoussaini M, Carroll JS, Jiang X, Finucane H, Adams M, Adank MA, Ahsan H, Aittomaki K, Anton-Culver H, Antonenkova NN, Arndt V, Aronson KJ, Arun B, Auer PL, Bacot F, Barrdahl M, Baynes C, Beckmann MW, Behrens S, Benitez J, Bermisheva M, Bernstein L, Blomqvist C, Bogdanova NV, Bojesen SE, Bonanni B, Borresen-Dale AL, Brand JS, Brauch H, Brennan P, Brenner H, Brinton L, Broberg P, Brock IW, Broeks A, Brooks-Wilson A, Brucker SY, Bruning T, Burwinkel B, Butterbach K, Cai Q, Cai H, Caldes T, Canzian F, Carracedo A, Carter BD, Castelao JE, Chan TL, David Cheng TY, Seng Chia K, Choi JY, Christiansen H, Clarke CL, Collaborators N, Collee M, Conroy DM, Cordina-Duverger E, Cornelissen S, Cox DG, Cox A, Cross SS, Cunningham JM, Czene K, Daly MB, Devilee P, Doheny KF, Dork T, Dos-Santos-Silva I, Dumont M, Durcan L, Dwek M, Eccles DM, Ekici AB, Eliassen AH, Ellberg C, Elvira M, Engel C, Eriksson M, Fasching PA, Figueroa J, Flesch-Janys D, Fletcher O, Flyger H, Fritschi L, Gaborieau V, Gabrielson M, Gago-Dominguez M, Gao YT, Gapstur SM, Garcia-Saenz JA, Gaudet MM, Georgoulias V, Giles GG, Glendon G, Goldberg MS, Goldgar DE, Gonzalez-Neira A, Grenaker Alnaes GI, Grip M, Gronwald J, Grundy A, Guenel P, Haeberle L, Hahnen E, Haiman CA, Hakansson N, Hamann U, Hamel N, Hankinson S, Harrington P, Hart SN, Hartikainen JM, Hartman M, Hein A, Heyworth J, Hicks B, Hillemanns P, Ho DN, Hollestelle A, Hooning MJ, Hoover RN, Hopper JL, Hou MF, Hsiung CN, Huang G, Humphreys K, Ishiguro J, Ito H, Iwasaki M, Iwata H, Jakubowska A, Janni W, John EM, Johnson N, Jones K, Jones M, Jukkola-Vuorinen A, Kaaks R, Kabisch M, Kaczmarek K, Kang D, Kasuga Y, Kerin MJ, Khan S, Khusnutdinova E, Kiiski JI, Kim SW, Knight JA, Kosma VM, Kristensen VN, Kruger U, Kwong A, Lambrechts D, Le Marchand L, Lee E, Lee MH, Lee JW, Neng Lee C, Lejbkowicz F, Li J, Lilyquist J, Lindblom A, Lissowska J, Lo WY, Loibl S, Long J, Lophatananon A, Lubinski J, Luccarini C, Lux MP, Ma ESK, MacInnis RJ, Maishman T, Makalic E, Malone KE, Kostovska IM, Mannermaa A, Manoukian S, Manson JE, Margolin S, Mariapun S, Martinez ME, Matsuo K, Mavroudis D, McKay J, McLean C, Meijers-Heijboer H, Meindl A, Menendez P, Menon U, Meyer J, Miao H, Miller N, Taib NAM, Muir K, Mulligan AM, Mulot C, Neuhausen SL, Nevanlinna H, Neven P, Nielsen SF, Noh DY, Nordestgaard BG, Norman A, Olopade OI, Olson JE, Olsson H, Olswold C, Orr N, Pankratz VS, Park SK, Park-Simon TW, Lloyd R, Perez JIA, Peterlongo P, Peto J, Phillips KA, Pinchev M, Plaseska-Karanfilska D, Prentice R, Presneau N, Prokofyeva D, Pugh E, Pylkas K, Rack B, Radice P, Rahman N, Rennert G, Rennert HS, Rhenius V, Romero A, Romm J, Ruddy KJ, Rudiger T, Rudolph A, Ruebner M, Rutgers EJT, Saloustros E, Sandler DP, Sangrajrang S, Sawyer EJ, Schmidt DF, Schmutzler RK, Schneeweiss A, Schoemaker MJ, Schumacher F, Schurmann P, Scott RJ, Scott C, Seal S, Seynaeve C, Shah M, Sharma P, Shen CY, Sheng G, Sherman ME, Shrubsole MJ, Shu XO, Smeets A, Sohn C, Southey MC, Spinelli JJ, Stegmaier C, Stewart-Brown S, Stone J, Stram DO, Surowy H, Swerdlow A, Tamimi R, Taylor JA, Tengstrom M, Teo SH, Beth Terry M, Tessier DC, Thanasitthichai S, Thone K, Tollenaar R, Tomlinson I, Tong L, Torres D, Truong T, Tseng CC, Tsugane S, Ulmer HU, Ursin G, Untch M, Vachon C, van Asperen CJ, Van Den Berg D, van den Ouweland AMW, van der Kolk L, van der Luijt RB, Vincent D, Vollenweider J, Waisfisz Q, Wang-Gohrke S, Weinberg CR, Wendt C, Whittemore AS, Wildiers H, Willett W, Winqvist R, Wolk A, Wu AH, Xia L, Yamaji T, Yang XR, Har Yip C, Yoo KY, Yu JC, Zheng W, Zheng Y, Zhu B, Ziogas A, Ziv E, Investigators A, ConFab AI, Lakhani SR, Antoniou AC, Droit A, Andrulis IL, Amos CI, Couch FJ, Pharoah PDP, Chang-Claude J, Hall P, Hunter DJ, Milne RL, Garcia-Closas M, Schmidt MK, Chanock SJ, Dunning AM, Edwards SL, Bader GD, Chenevix-Trench G, Simard J, Kraft P, Easton DF (2017) Association analysis identifies 65 new breast cancer risk loci. Nature 551 (7678):92-94. doi:10.1038/nature24284
15. Blanche P, Dartigues JF, Jacqmin-Gadda H (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 32 (30):5381-5397. doi:10.1002/sim.5958
16. Blanche P, Kattan MW, Gerds TA (2018) The c-index is not proper for the evaluation of t-year predicted risks. Biostatistics. doi:10.1093/biostatistics/kxy006
17. Pfeiffer RM, Park Y, Kreimer AR, Lacey JV, Jr., Pee D, Greenlee RT, Buys SS, Hollenbeck A, Rosner B, Gail MH, Hartge P (2013) Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. PLoS Med 10 (7):e1001492. doi:10.1371/journal.pmed.1001492
18. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 74:167-176. doi:10.1016/j.jclinepi.2015.12.005
19. Collins GS, Ogundimu EO, Altman DG (2016) Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 35 (2):214-226. doi:10.1002/sim.6787
20. Team RDC (2017) A language and Environment for Statistical Computing. R: Foundation for Statistical Computing
21. Akdeniz D, Schmidt MK, Seynaeve CM, McCool D, Giardiello D, van den Broek AJ, Hauptmann M, Steyerberg EW, Hooning MJ (2018) Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. Breast 44:1-14. doi:10.1016/j.breast.2018.11.005
22. Armstrong N, Ryder S, Forbes C, Ross J, Quek RG (2019) A systematic review of the international prevalence of BRCA mutation in breast cancer. Clin Epidemiol 11:543-561. doi:10.2147/CLEP.S206949
23. Bueno-de-Mesquita JM, Nuyten DS, Wesseling J, van Tinteren H, Linn SC, van de Vijver MJ (2010) The impact of inter-observer variation in pathological assessment of node-negative breast cancer on clinical risk assessment and patient selection for adjuvant systemic treatment. Ann Oncol 21 (1):40-47. doi:10.1093/annonc/mdp273
24. Whittle R, Peat G, Belcher J, Collins GS, Riley RD (2018) Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. J Clin Epidemiol 102:38-49. doi:10.1016/j.jclinepi.2018.05.008
25. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M (2019) Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. Stat Med 38 (18):3444-3459. doi:10.1002/sim.8183
26. Pflieger LT, Mason CC, Facelli JC (2017) Uncertainty quantification in breast cancer risk prediction models using self-reported family health history. J Clin Transl Sci 1 (1):53-59. doi:10.1017/cts.2016.9
27. Groenwold RH, White IR, Donders AR, Carpenter JR, Altman DG, Moons KG (2012) Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. CMAJ 184 (11):1265-1269. doi:10.1503/cmaj.110977

28. Janssen KJ, Donders AR, Harrell FE, Jr., Vergouwe Y, Chen Q, Grobbee DE, Moons KG (2010) Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 63 (7):721-727. doi:10.1016/j.jclinepi.2009.12.008
29. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Jr., Chen Q, Grobbee DE, Moons KG (2009) Dealing with missing predictor values when applying clinical prediction models. Clin Chem 55 (5):994-1001. doi:10.1373/clinchem.2008.115345
30. Royston P, Altman DG (2013) External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 13:33. doi:10.1186/1471-2288-13-33
31. van Houwelingen HC (2000) Validation, calibration, revision and combination of prognostic survival models. Stat Med 19 (24):3401-3415
32. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH (2019) How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. J Clin Epidemiol 105:136-141. doi:10.1016/j.jclinepi.2018.09.001
33. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110:12-22. doi:10.1016/j.jclinepi.2019.02.004
34. Ming C, Viassolo V, Probst-Hensch N, Chappuis PO, Dinov ID, Katapodi MC (2019) Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res 21 (1):75. doi:10.1186/s13058-019-1158-4
35. Torkamani A, Wineinger NE, Topol EJ (2018) The personal and clinical utility of polygenic risk scores. Nat Rev Genet 19 (9):581-590. doi:10.1038/s41576-018-0018-x
36. Mellemkjaer L, Dahl C, Olsen JH, Bertelsen L, Guldberg P, Christensen J, Borresen-Dale AL, Stovall M, Langholz B, Bernstein L, Lynch CF, Malone KE, Haile RW, Andersson M, Thomas DC, Concannon P, Capanu M, Boice JD, Jr., Group WSC, Bernstein JL (2008) Risk for contralateral breast cancer among carriers of the CHEK2*1100delC mutation in the WECARE Study. Br J Cancer 98 (4):728-733. doi:10.1038/sj.bjc.6604228

SUPPLEMENTARY MATERIAL

1. From Manchester guidelines to Manchester formula

A simple formula has been proposed by Basu et al. to calculate the life-time risk of contralateral breast cancer (CBC) based on a literature review of risk factors. We translated the life-time risk to 5- and 10-year CBC risk to compare the prediction performance with CBCrisk and PredictCBC. A complete overview is given in the following schema:

Manchester guidelines (life-time CBC risk)	Manchester formula (CBC risk at 5 and 10 years)
Step 1: calculate the number of years of CBC risk using 80 years as the average life expectancy. The number of years of CBC risk (N) can be calculated as:	**Step 1:** set CBC risk at 5 and 10 years assuming the number of years at risk of CBC (N) be 5 and 10, respectively:
N = 80 years – patient age (years)	N = 5 or 10
Step 2: calculate the life-time risk (L) using the annual incidence of CBC at 0.5% among patients diagnosed with first invasive breast cancer:	**Step 2:** calculate the CBC risk at 5 and 10 years using an annual incidence of CBC of 0.5% among patients diagnosed with first invasive breast cancer:
L = N × 0.5%	L = 5 × 0.5% (5-year CBC risk) L = 10 × 0.5% (10-year CBC risk)
Step 3: The life-time risk can be modified based on patient's personal risk profile as: • For patients with estrogen-receptor positive disease: L × 0.5; • For patients with a *BRCA1/2* germline mutation*: L × 4; • For patients with oophorectomy under age of 40 years: L × 0.5 • For patients with a family history of breast cancer*: L × 2; * For patients with a *BRCA1/2* germline mutation and a family history: L × 4.	**Step 3:** according to the information of the risk factors available, calculate the 5- and 10-year CBC personalized risk (Y): $Y = L \times (1-0.5X_1) \times (1-0.5X_2) \times (1+X_3+3X_4 - X_5)]$ Where: X_1 = 1 (if estrogen-receptor positive, 0 else) X_2 = 1 (if oophorectomy under 40 years, 0 else) X_3 = 1 (if family history, 0 else) X_4 = 1 (if BRCA mutation, 0 else) X5 = 1 (if BRCA mutation and family history, 0 else)

2. Sensitivity analysis

The PredictCBC and CBCrisk models considered contralateral breast cancer (CBC) more than 3 and 6 months after the first invasive breast cancer (BC) diagnosis as metachronous, respectively. A sensitivity analysis was done to evaluate the performance of CBCrisk by

considering metachronous CBC after 6 months since first BC. Discrimination accuracy at 5 years was 0.59 with the corresponding 95% confidence interval (CI) between 0.57 and 0.61 (**Supplementary Figure 3**), and the corresponding 95% prediction interval (PI) between 0.54 and 0.64. Discrimination accuracy at 10 years was 0.58 (95% CI: 0.57–0.59; 95% PI: 0.55–0.61, **Supplementary Figure 3**). The expected/observed (E/O) ratio at 5 years was 0.89 (95% CI:0.53–1.49.; 95% PI: 0.21–3.72) and at 10 years was 0.84 (95% CI: 0.53–1.34; 95% PI: 0.23–3.13) (**Supplementary Table 1)**. The calibration slope was 1.26 (95% CI: 1.01–1.51; 95% PI: 1.01–1.51). An overview of prediction performances of the main results and those from the sensitivity analysis are shown in **Supplementary Table 2**.

Supplementary Table 1: Calibration performances of CBCrisk from the sensitivity analysis defining CBC as an event 6 months after the first breast cancer diagnosis

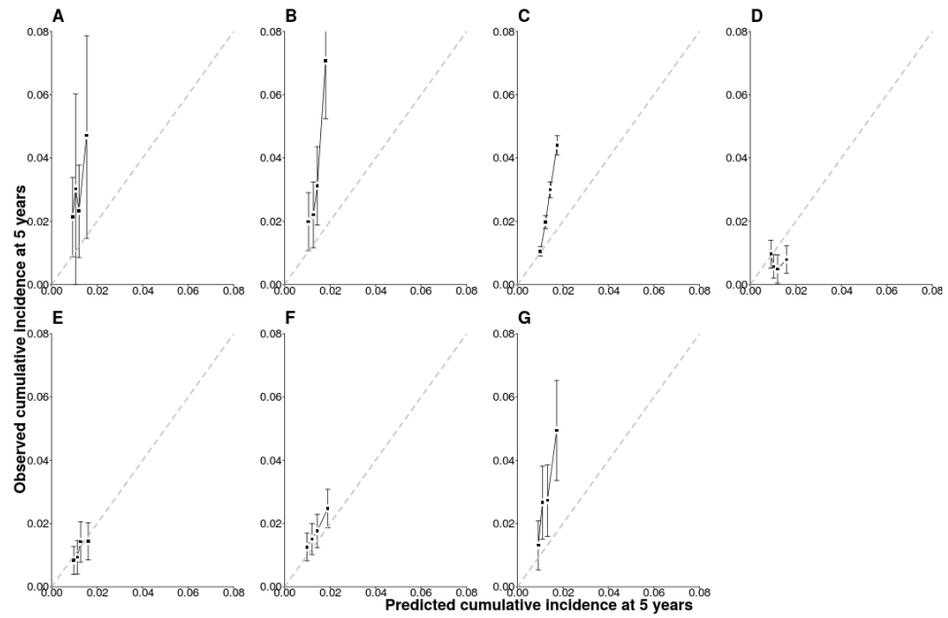
Validation dataset	E/O ratio at 5 years (95% CI)	E/O ratio at 10 years (95% CI)	Calibration slope (95% CI)
Europe - Other	0.92 (0.80 - 1.04)	0.76 (0.69 - 0.83)	1.17 (0.44 - 1.91)
Europe - Scandinavia	1.61 (1.29 - 1.93)	1.24 (1.08 - 1.39)	0.84 (0.14 -1.55)
Europe - UK	1.77 (1.38 - 2.17)	1.82 (1.53 - 2.11)	0.85 (-0.03 - 1.73)
Netherlands - BOSOM	0.47 (0.38 - 0.56)	0.51 (0.44 - 0.58)	1.33 (0.73 - 1.92)
Netherlands - EMC	0.49 (0.39 - 0.59)	0.44 (0.38 - 0.50)	1.22 (0.68 - 1.77)
Netherlands - NCR	0.58 (0.55 - 0.61)	0.55 (0.53 - 0.56)	1.40 (1.11 - 1.69)
US and Australia	0.56 (0.40 - 0.72)	0.66 (0.52 - 0.80)	1.15 (0.27 - 2.03)
Meta-analysis	0.89 (0.53 - 1.49)	0.84 (0.53 - 1.34)	1.26 (1.01 - 1.51)
95% PI	0.21 - 3.72	0.23 - 3.13	1.01 - 1.51

Abbreviations: E/O: expected/observed; PI: prediction interval;
BOSOM: Breast Cancer Outcome Study of Mutation carriers; EMC: Erasmus Medical Center; UK: United Kingdom; NCR: Netherlands Cancer Registry

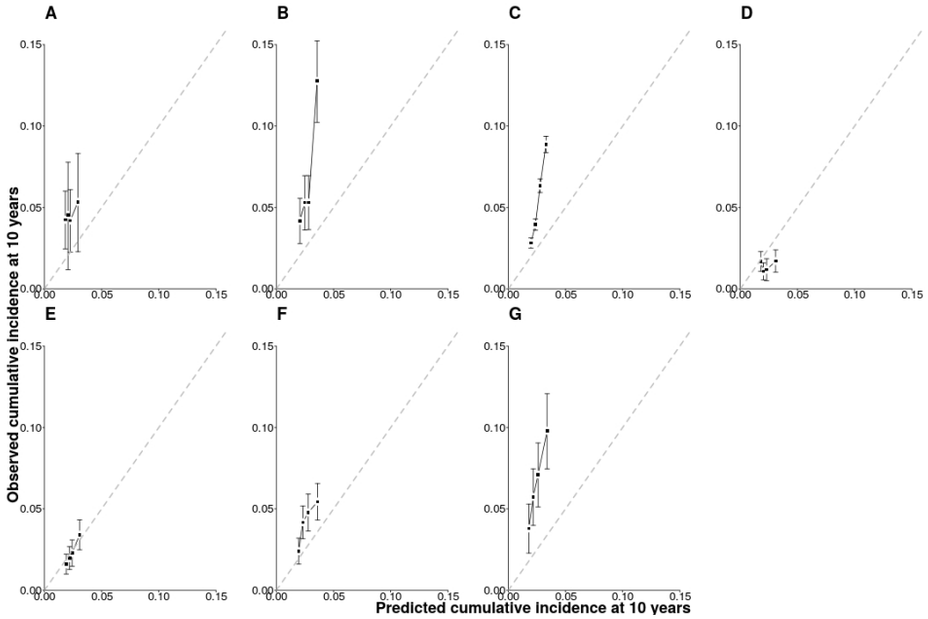
Supplementary Table 2: An overview of prediction performances of the CBCrisk model

Characteristics	CBCrisk CBC definition 3 months	CBCrisk CBC definition 6 months (sensitivity analysis)
Discrimination		
AUC at 5 years (95% PI)	0.59 (0.54–0.64)	0.59 (0.54–0.64)
AUC at 10 years (95% PI)	0.58 (0.55–0.61)	0.58 (0.55–0.61)
Calibration		
E/O ratio at 5 years (95% PI)	0.86 (0.20–3.75)	0.89 (0.21–3.72)
E/O ratio at 10 years (95% PI)	0.82 (0.27–2.51)	0.84 (0.23–3.13)
Slope (95% PI)	1.26 (1.01–1.50)	1.26 (1.01–1.51)

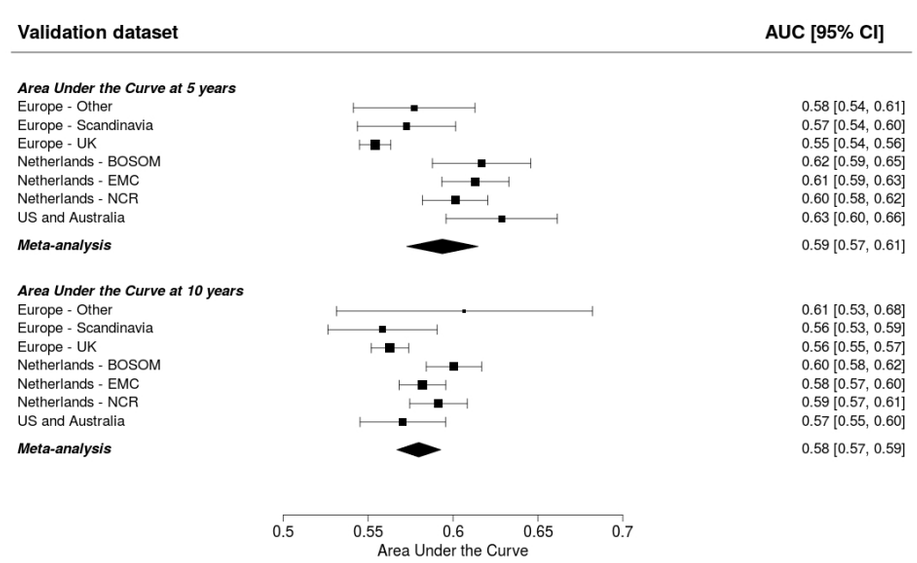
Abbreviations:
AUC: Area-Under-the-Curve;
PI: prediction interval
E/O: expected/observed



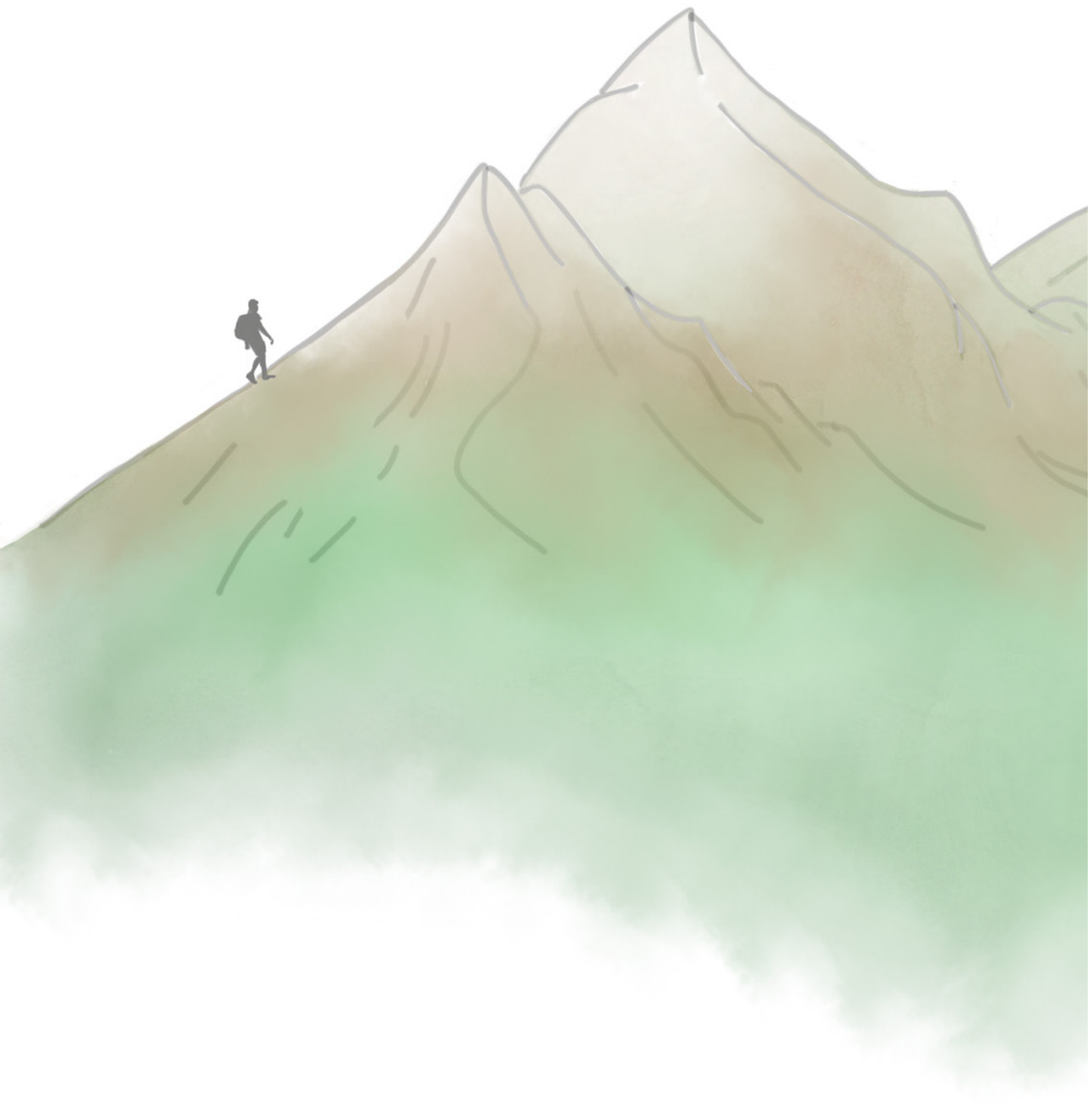
Supplementary Figure 1: Calibration plot of CBCrisk at 5 years
The x-axis represents the predicted cumulative incidence of contralateral breast cancer at 5 years and the y-axis the observed cumulative incidence at 5 years. The black dots indicate the calibration for quartiles of predicted values. Vertical black bars indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Netherlands - BOSOM; Panel B: Netherlands - EMC; Panel C: Netherlands - NCR; Panel D: Europe – Scandinavia; Panel E: United States and Australia; Panel F: Europe – Other; Panel G: Europe – United Kingdom.



Supplementary Figure 2: Calibration plot of CBCrisk at 10 years
The x-axis represents the predicted cumulative incidence of contralateral breast cancer at 10 years and the y-axis the observed cumulative incidence at 10 years. The black dots indicate the calibration for quartiles of predicted values. Vertical black bars indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Netherlands - BOSOM; Panel B: Netherlands - EMC; Panel C: Netherlands - NCR; Panel D: Europe – Scandinavia; Panel E: United States and Australia; Panel F: Europe – Other; Panel G: Europe – United Kingdom.



Supplementary Figure 3: Sensitivity analysis of CBCrisk prediction performances. A metachronous contralateral breast cancer was defined after 6 months since the first breast cancer diagnosis. The upper and lower panel show the discrimination assessed by a time-dependent Area-Under-the-Curve at 5 and 10 years, respectively. The black squares for each dataset indicate the estimated accuracy of a model built on all remaining studies or geographic areas. The black horizontal lines indicate the corresponding 95% confidence intervals of the estimated accuracy (interval whiskers). The black diamonds indicate the mean with the corresponding 95% confidence interval of the predictive accuracy.



Chapter 4

PredictCBC-2.0: a contralateral breast cancer risk prediction model developed and validated in ~200,000 patients

Submitted for publication
Preprint available here: https://www.researchsquare.com/article/rs-1767532/v1.

Daniele Giardiello
Maartje J. Hooning
Michael Hauptmann
Renske Keeman, B. A. M. Heemskerk-Gerritsen, Heiko Becher, Carl Blomqvist, Stig E. Bojesen, Manjeet K. Bolla, Nicola J. Camp, Kamila Czene, Peter Devilee Diana M. Eccles, Peter A. Fasching, Jonine D. Figueroa, Henrik Flyger Montserrat García-Closas, Christopher A. Haiman, Ute Hamann, John L. Hopper, Anna Jakubowska, Flora E. Leeuwen, Annika Lindblom, Jan Lubiński, Sara Margolin, Maria Elena Martinez, Heli Nevanlinna, Ines Nevelsteen, Saskia Pelders, Paul D.P. Pharoah, Sabine Siesling, Melissa C. Southey, Annemieke H. van der Hout, Liselotte P. van Hest, Jenny Chang-Claude, Per Hall, Douglas F. Easton,
Ewout W. Steyerberg
Marjanka K. Schmidt

ABSTRACT

Background

Prediction of contralateral breast cancer (CBC) risk is challenging due to moderate performances of the known risk factors. We aimed to improve our previous risk prediction model (PredictCBC) by updated follow-up and including additional risk factors.

Methods

We included data from 207,510 invasive breast cancer patients participating in 23 studies. 8,225 CBC events occurred over a median follow-up of 10.2 years. In addition to the previously included risk factors, PredictCBC-2.0 included *CHEK2* c.1100delC, a 313 variant polygenic risk score (PRS-313), body mass index (BMI), and parity. Fine and Gray regression was used to fit the model. Calibration and a time-dependent Area Under the Curve (AUC) at 5 and 10 years were assessed to determine the performance of the models. Decision curve analysis was performed to evaluate the net benefit of PredictCBC-2.0 and previous PredictCBC models.

Results

The discrimination of PredictCBC-2.0 at 10 years was higher than PredictCBC with an AUC of 0.65 (95% prediction intervals (PI):0.56–0.74) versus 0.63 (95%PI:0.54–0.71). PredictCBC-2.0 was well-calibrated with an observed/expected (O/E) ratio at 10 years of 0.92 (95%PI:0.34–2.54). Decision curve analysis for contralateral preventive mastectomy (CPM) showed potential clinical utility of PredictCBC-2.0 between thresholds of 4–12% 10-year CBC risk for *BRCA1/2* mutation carriers and non-carriers.

Conclusions

Additional genetic information beyond *BRCA1/2* germline mutations improved CBC risk prediction and might help tailor clinical decision making towards CPM or alternative preventive strategies. Identifying patients who benefit from CPM, especially in the general breast cancer population, remains challenging.

INTRODUCTION

Contralateral breast cancer (CBC) is the most common second primary cancer among women diagnosed with first primary invasive breast cancer (BC)[1]. CBC accounts for approximately 40-50% of all new secondary cancers in women with first primary invasive BC and has potentially less favorable prognosis[2-6]. Worries regarding CBC risk have increased the demand for contralateral preventive mastectomy (CPM)[7, 8]. However, the impact of CPM on survival is uncertain, especially in women with low risk to develop a CBC[9-13].Thus, improved CBC risk prediction is important in order to inform decision making on surveillance and preventive strategies. Currently, the most important factor for decision making on CPM is the *BRCA1/2* mutations status[14].

We previously developed and cross-validated two models using data from 132,756 invasive BC patients with a median follow-up of 8.8 years including 4,672 CBC events[15]. One model (PredictCBC-1A) was developed including information about *BRCA1/2* mutation status and another (PredictCBC-1B) for the general breast cancer population of genetically untested women. Two other specific CBC prediction tools are currently available in the literature: the Manchester formula (part of the Manchester guidelines for CPM) and CBCrisk[15-18].

In addition to *BRCA1/2* mutations, other genetic risk factors for breast cancer are also associated with CBC risk. In particular, there is substantial evidence that the *CHEK2* c.1100delC variant increases the risk of developing CBC[19, 20]. In addition, polygenic risk scores (PRS) of common variants, developed for association with a first breast cancer have been shown to predict CBC in the general BC population and in *BRCA1/2* mutation carriers[21-24], particularly the extensively validated 313 SNP PRS[25]. With regard to the lifestyle and reproductive factors, there is evidence that body mass index (BMI) and parity at or around the time of the first primary invasive BC diagnosis are associated with CBC risk[26].

Our aim was to refit PredictCBC models incorporating these additional risk factors. We utilized the same dataset but with updated follow-up, and added additional studies, especially one large study of *BRCA1* and *BRCA2* mutation carriers. We evaluated the potential improvement in prediction performance and utility for clinical decision making of the updated models for both *BRCA1/2* carriers as the general (non-tested) breast cancer population (PredictCBC-2.0).

MATERIAL AND METHODS

Study population and available data

We used the data from the same five main sources previously used to develop PredictCBC models to develop the PredictCBC-2.0 models including updated follow-up information, additional patients and CBC events[15]. Two studies were addionally included from the Breast Cancer Association Consortium (BCAC) compared to the version of the BCAC data used to develop PredictCBC-1A and PredictCBC-1B models. Most of the studies were either population- or hospital-based series; and most women were of European-descent (**Supplementary Tables 1-2,** available online). We also additionally included patients selected from the Hereditary Breast and Ovarian cancer study in the Netherlands (HEBON)[27], a nationwide study based on clinical genetic centers. The eligibility criteria were the same as previously: briefly, we included female patients with invasive first primary BC with no sign of distant metastases at diagnosis or prior history of cancer (except for non-melanoma skin cancer)[15]. We included women diagnosed after 1990 so that diagnostic and treatment procedures were close to modern practice while follow-up was sufficient to study CBC incidence. In total 207,510 women from 23 studies were included. All studies were approved by the appropriate ethics and scientific review boards. All women provided written informed consent; or, for some Dutch cohorts as applicable, the secondary use of clinical data was in accordance with Dutch legislation and codes of conduct[28, 29]. Information on the factors included in the analyses, follow-up per dataset, and study design are in **Supplementary Table 2,** available online.

Statistical analyses

Primary endpoint and follow-up

The primary endpoint in the analyses was incidence of invasive or in situ metachronous CBC. Follow-up started 3 months after invasive first primary BC diagnosis, to exclude synchronous CBCs, and ended at date of CBC, distant metastasis (but not a loco-regional relapse), CPM, or last date of follow-up (due to death, loss to follow-up, or end of study), whichever occurred first. For 36,553 (17.6%) women, from BCAC and HEBON, recruitment or blood sampling for DNA testing occurred more than 3 months after diagnosis of the first primary BC. For these women, follow-up (prevalent cases), started at recruitment or at the date of blood draw or at DNA test result (left truncation). Patients who underwent CPM during the follow-up were censored because of negligible CBC risk after a CPM[30]. Missing data were multiply imputed by chained equations (MICE) to avoid loss of information due to case-wise deletion[31-33](**Supplementary Material**, available online).

Model development and validation

We used multivariable Fine and Gray regression models to account for death and distant metastases as competing events[34]. Analyses were stratified by study to allow baseline hazard (sub)distributions to differ across studies. The assumption of proportional subdistribution hazards was graphically checked using Schoenfeld residuals[35]. The resulting subdistribution hazard ratios (sHRs) and corresponding 95% confidence intervals (CI) were pooled from 5 imputed data sets using Rubin's rules[33]. We re-estimated the coefficients of PredictCBC-1A and PredictCBC-1B, and we re-fitted the PredictCBC models using the extended data set with updated follow-up time. PredictCBC-1A, developed including information about *BRCA1/2* mutation carrier status, was extended by including *CHEK2* c.1110delC status, PRS-313, BMI, and parity (hereafter: PredictCBC-2.0A)[15]. *CHEK2* c.1110delC and PRS-313 were derived from the BCAC database, as published previously[25, 36, 37]. We extended PredictCBC-1B, developed for genetically untested women, incorporating BMI and parity (hereafter: PredictCBC-2.0B). Potential non-linear relations between continuous predictors and CBC risk were investigated using restricted cubic splines with three knots.

The validity of the model was investigated by leave-one-study-out cross-validation[38]. In each validation cycle, all studies were analyzed except one, in which the validity of the model was evaluated. Since some BCAC studies had insufficient CBC events required for reliable validation, we used the geographic area as unit for splitting[38-40]. Nineteen out of 23 studies were combined in 4 geographic areas. (**Supplementary Table 3**, available online). A total of 8 units of splitting including 4 geographic ares and 4 studies were used to cross-validated the models.

The performance of the PredictCBC-2.0 was assessed by discrimination, i.e., the ability to differentiate between patients diagnosed with CBC and those who were not, and by calibration, which measures the agreement between the actual (observed) risk and CBC risk estimated by the prediction models (predicted). Discrimination was quantified by time-dependent areas under the ROC curve (AUCs) based on Inverse Censoring Probability Weighting at 5 and 10 years[41]. Values of AUCs close to 1 indicate good discrimination, while values close to 0.5 indicated poor discrimination. Calibration was assessed by the observed to expected (O/E) ratio and calibration plots at 5 and 10 years[42, 43]. An O/E ratio lower or higher than 1 indicates that average predictions are too high or low, respectively.

To consider heterogeneity among studies, a random-effect meta-analysis was performed to provide summaries of discrimination and calibration performance. The 95% prediction intervals (PI) indicate the likely performance of the model in a new dataset. The summary performances of PredictCBC-2.0 and 1.0 models were compared to evaluate whether adding the new predictors improved the performance of CBC risk prediction. We developed and validated the risk prediction model following the

Transparent Reporting of a Multivariable Prediction model for Individual Prognosis or Diagnosis (TRIPOD)statement[44]. Analyses were done in SAS (SAS Institute Inc., Cary, NC, USA) and R (version 3.6.1).

Clinical utility

The clinical utility of the prediction models was evaluated using decision curve analysis (DCA)[45, 46]. A key metric DCA is the net benefit, which is the number of true-positive classifications (in this example: the number of CPMs in patients who would have developed a CBC) minus the weighted number of false-positive classifications (in this example: the number of unnecessary CPMs in patients who would not have developed a CBC). The false positives are weighted by a factor related to the relative harm of a missed CBC versus an unnecessary CPM. The weighting is derived from the threshold probability to develop a CBC using a fixed time horizon (e.g., CBC risk at 5 or 10 years) [47]. For example, a threshold of 10% implies that CPM in 10 patients, of whom one would develop CBC if untreated, is acceptable (thus performing 9 unnecessary CPMs). The net benefit of a prediction model is traditionally compared with the strategies of treat all or treat none. Since the use of CPM is generally only considered among *BRCA1/2* mutation carriers, the decision curve analysis was reported among *BRCA1/2* mutation carriers and non-carriers separately[48]. Among patients not tested for *BRCA1/2* germline mutations, we assumed that the decision for CPM is based on family history of breast cancer. Net benefits of PredictCBC-2.0A and PredictCBC-2.0B were compared with net benefit of PredictCBC-1A and 1B, respectively, to assess the potential improvement in clinical utility of the updated models.

RESULTS

A total of 207,510 women with invasive first primary BC diagnosed between 1990 and 2017, with 8,225 CBC events (6,828 invasive, 1,397 in situ), from 23 studies, were used for prediction modeling for CBC risk (**Supplementary Table 2**, available online). Median follow-up time was 10.2 years and CBC cumulative incidences at 5 and 10 years were 2.2% and 4.1%, respectively. Details of the studies and patient, tumor, and treatment characteristics are provided in **Supplementary Table 4** (available online). The multivariable models with estimates for all included factors are shown in **Table 1**.

Most of factors were independently associated with CBC risk, including the new factors incorporated in the PredictCBC-2.0 models, i.e., s BMI, parity, *CHEK2* c.1110delC, and PRS-313. There was no evidence against log-linear relationships between BMI, parity and PRS-313 and CBC risk. Non-linearity between age at first BC diagnosis and CBC risk was accounted for with a linear spline at age 60 years. The formulae of the PredictCBC models are provided in **Supplementary Methods** (available online). To calculate the predicted CBC cumulative incidence, we used the event-free baseline probability of the Netherlands Cancer Registry (NCR), as previously[15].

Table 1. Multivariable subdistribution hazard models for contralateral breast cancer risk

Factor (reference)		PredictCBC-2.0A	PredictCBC-2.0B
		sHR (95% CI)	sHR (95% CI)
Age at PBC, *years* (75th vs 25th quartile: 66 vs 48)		0.87[a] (0.83 - 0.90)	0.82[a] (0.78 - 0.85)
Body mass index, *kg/m*^{2} (75th vs 25th quartile: 28.4 vs 22.7)		1.06 (1.03 - 1.09)	1.06 (1.03 - 1.09)
Parity (75th vs 25th quartile: 3 vs 1)		0.85 (0.82 - 0.88)	0.86 (0.83 - 0.90)
Family history (yes)		1.17 (1.12 - 1.23)	1.35 (1.29 - 1.42)
BRCA mutation			
	BRCA1 vs non-carrier	4.79 (4.43 - 5.17)	-
	BRCA2 vs non-carrier	3.09 (2.72 - 4.25)	-
PRS_{313}[b] (75th vs 25th quartile: -0.49 vs 0.32)		1.35 (1.31 - 1.39)	-
CHEK2 c.1100delC mutation (present)		2.75 (2.85 - 3.34)	-
Nodal status of FBC (positive)		0.99 (0.93 - 1.05)	0.99 (0.93 - 1.04)
Tumor size category of FBC, *cm*			
	(2,5] vs ≤ 2	0.99 (0.94 - 1.05)	1.01 (0.96 - 1.07)
	> 5 vs ≤ 2	1.23 (1.10 - 1.36)	1.22 (1.09 - 1.36)
Morphology of FBC (lobular including mixed)		1.19 (1.12 - 1.27)	1.17 (1.10 - 1.24)
Grade of FBC			
	Moderately differentiated vs well differentiated (II vs I)	0.93 (0.88 - 0.99)	0.98 (0.93 - 1.04)
	Poorly differentiated vs well differentiated (III vs I)	0.85 (0.79 - 0.91)	0.95 (0.88 - 1.01)
Chemotherapy (yes)		0.75 (0.70 - 0.80)	0.75 (0.70 - 0.80)
Radiotherapy to the breast (yes)		0.93 (0.89 - 0.98)	0.95 (0.90 - 0.99)
ER with endocrine therapy			
	negative/no vs positive/yes	1.53 (1.43 - 1.65)	1.78 (1.67 - 1.90)
	positive/no vs positive/yes	1.95 (1.83 - 2.07)	1.94 (1.82 - 2.06)
HER2 with trastuzumab therapy			
	negative/no vs positive/yes	1.22 (1.09 - 1.38)	1.30 (1.15 - 1.46)
	positive/no vs positive/yes	1.12 (0.97 - 1.28)	1.14 (1.00 - 1.31)

Abbreviations:
vs: versus; sHR: subdistributional hazard ratio; CI: confidence interval; PRS: polygenic risk score; PBC: first primary breast cancer; ER: estrogen receptor; HER2: human epidermal growth factor 2;
[a] age was parametrized as a linear spline with one interior knot at 60 years. For representation purposes, we here provide the sHR for the 75th versus the 25th percentile.
[b] PRS standardized by the same standard deviation (SD) used by Mavaddat et al (SD=0.61)[25].

The AUCs at 5 and 10 years of PredictCBC-2.0A were higher than of PredictCBC-1A at 5 years: 0.66, 95% prediction interval (PI): 0.55–0.76 versus 0.62 (95%PI:0.51–0.74); and at 10 years: 0.65 (95%PI:0.56–0.74) versus 0.63 (95%PI:0.54–0.71)(**Figure 1-2, Table 2**). The AUCs for PredictCBC-2.0B and PredictCBC-1B were both 0.59 (95%PI: PredictCBC-2.0B:0.51–0.68; PredictCBC-1B:0.49–0.69) at 5 years and both 0.58 (95%PI:0.51–0.65) at 10 years (**Figure 1-2, Table 2**).

The O/E ratio at 5 and 10 years across all versions of PredictCBC models ranged between 0.90 and 0.92 with similar 95%PIs (**Figure 1-2, Table 2**). Calibration plots of PredictCBC 2.0 models are provided in the **Supplementary Figures 1-4** (available online).

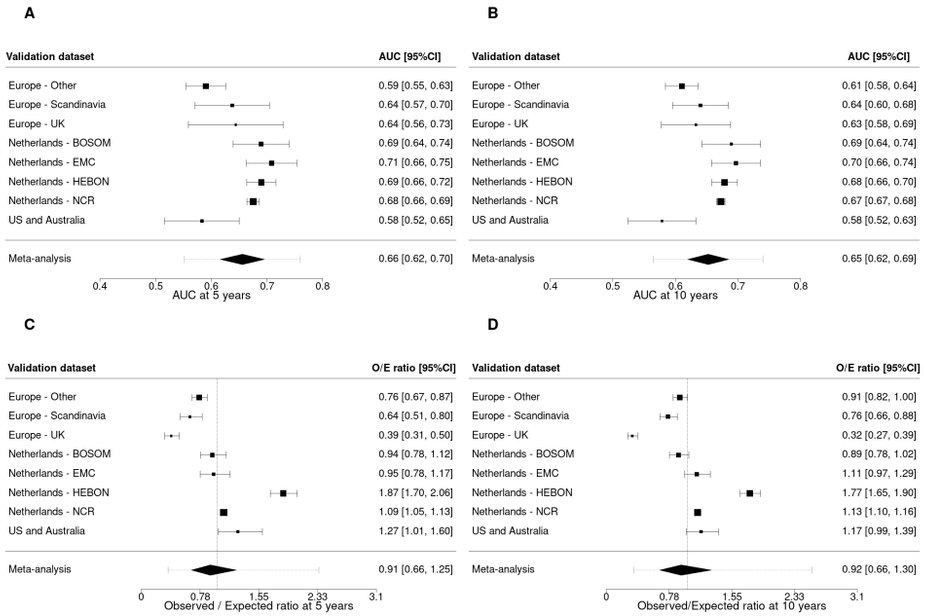


Figure 1. Analysis of predictive performance of PredictCBC-2.0A in leave-one-study-out cross-validation. Discrimination was assessed by a time-dependent AUC at 5 and 10 years (panel A and B, respectively). Calibration accuracy was measured with observed/expected (O/E) ratio at 5 and 10 years (panel C and D, respectively). The black squares indicate the estimated accuracy of a model built using all remaining studies or geographic areas. The black horizontal lines indicate the corresponding 95% confidence intervals of the estimated accuracy (interval whiskers). The black diamonds indicate the mean with the corresponding 95% confidence intervals of the predictive accuracy, and the dashed horizontal lines indicate the corresponding 95% prediction intervals.

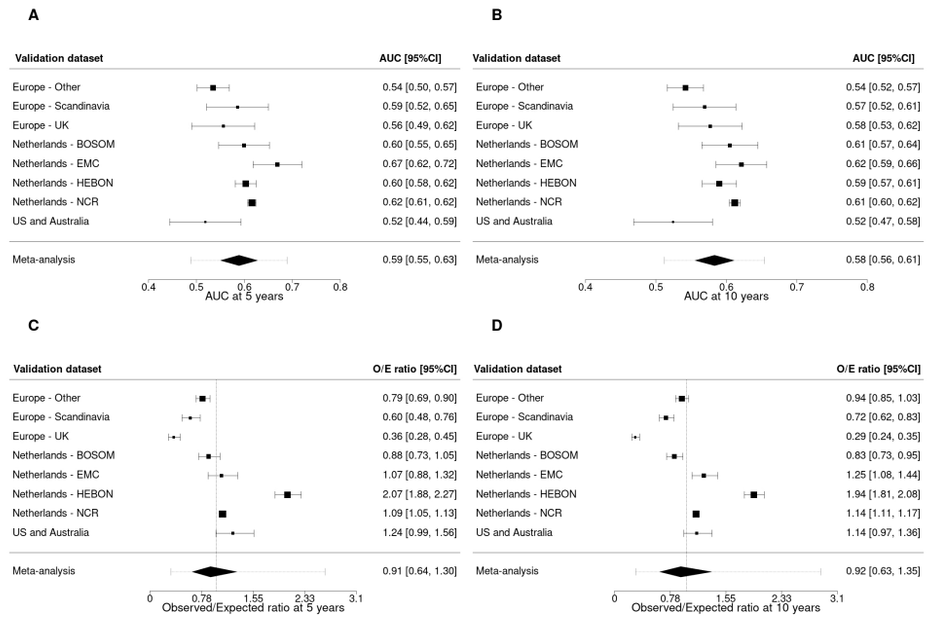


Figure 2. Analysis of predictive performance of PredictCBC-2.0B in leave-one-study-out cross-validation. Discrimination was assessed by a time-dependent AUC at 5 and 10 years (panel A and B, respectively). Calibration accuracy was measured with observed/expected (O/E) ratio at 5 and 10 years (panel C and D, respectively). The black squares indicate the estimated accuracy of a model built using all remaining studies or geographic areas. The black horizontal lines indicate the corresponding 95% confidence intervals of the estimated accuracy (interval whiskers). The black diamonds indicate the mean with the corresponding 95% confidence intervals of the predictive accuracy, and the dashed horizontal lines indicate the corresponding 95% prediction intervals.

Table 2. Summary of prediction performance of PredictCBC-1A, PredictCBC-1B, PredictCBC-2.0A and PredictCBC-2.0B with the corresponding 95% prediction intervals (PI) based on a leave-one-study out cross-validation procedure.

CBC risk prediction model	Performance measure			
	Discrimination		Calibration	
	AUC (95% PI)		O/E ratio (95% PI)	
	5-year	10-year	5-year	10-year
PredictCBC-1A	0.62 (0.51-0.74)	0.63 (0.54-0.71)	0.90 (0.36-2.24)	0.91 (0.34-2.48)
PredictCBC-2.0A	0.66 (0.55-0.76)	0.65 (0.56-0.74)	0.91 (0.35-2.34)	0.92 (0.34-2.54)
PredictCBC-1B	0.59 (0.49-0.69)	0.58 (0.51-0.65)	0.91 (0.32-2.55)	0.92 (0.30-2.80)
PredictCBC-2.0B	0.59 (0.51-0.68)	0.58 (0.51-0.65)	0.91 (0.31-2.63)	0.92 (0.30-2.87)

Abbreviations: AUC: Area under the Curve; CBC: contralateral breast cancer; PI: prediction interval; O/E = observed/expected

The decision curves showed the net benefit for a range of harm-benefit thresholds at 10-year CBC risk (**Figure 4**). We evalutated the potential clinical utility of PredictCBC-2A versus PredictCBC-1.0A for decision thresholds between 4-12% for the 10-year CBC risk among *BRCA1/2* mutation carriers and non-carriers (**Table 3**). For example, if consensus guidelines would indicate acceptability of one in 10 patients for whom a CPM is recommended developing CBC, a risk threshold of 10% may be used to define high and low risk *BRCA1/2* mutation carriers based on the absolute 10-year CBC risk prediction estimated by the models. Compared with a strategy recommending CPM to all *BRCA1*/2 mutation carriers, PredictCBC-1A avoids 76.9 net CPMs per 1,000 patients (**Table 3**). An additional 50.0 CPMs may be avoided using PredictCBC-2.0A compared to PredictCBC-1A. In contrast, almost no non-*BRCA1/2* mutation carriers had predictions above the 10% threshold (general BC population, **Table 3**); three necessary CPMs per 1,000 patients would be indicated using PredictCBC-2.0A. Analyses for PredictCBC-1B and PredictCBC-2.0B at 10 years suggested a potential clinical utility between 4-6% 10-year CBC risk for patients with and without family history (**Table 3 and Figure 4)**. No remarkable improvement in net benefit was detected using PredictCBC-2.0B compared to PredictCBC-1B in decision making regarding CPM (**Table 3 and Figure 4)**. Decision curves for CBC risk using PredictCBC and PredictCBC-2.0 at 5 years and the corresponding clinical utility showed similar patterns (**Supplementary Figures S5-6** and **Supplementary Table 5**, available online).

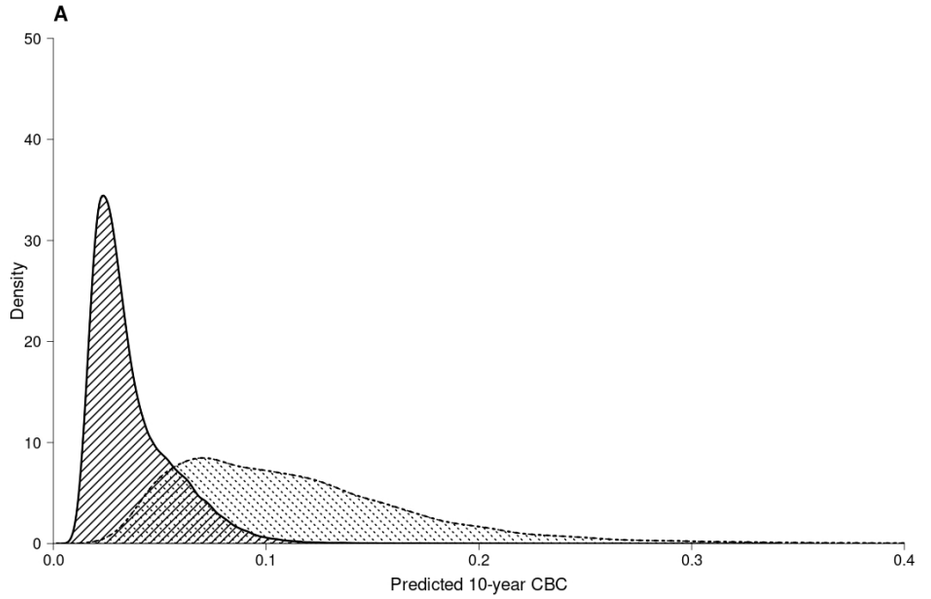


Figure 3. Density distribution of 10-year predicted contralateral breast cancer using PredictCBC version 2 models. **a** Density distribution of 10-year predicted contralateral breast cancer absolute risk using PredictCBC-2.0A within non-carriers (area with black solid lines) and *BRCA1/2* mutation carriers (area with black dashed lines). **b** Density distribution of 10-year predicted contralateral breast cancer absolute risk using PredictCBC-2.0B within patients without (first degree) family history (area with black solid lines) and patients with (first degree) family history (area with black dashed lines).

Table 3: Clinical utility of the 10-year contralateral breast cancer risk prediction models (PredictCBC-1A with PredictCBC-2.0A and PredictCBC-1B with PredictCBC-2.0B). For PredictCBC versions 1A and 2.0A, at the same probability threshold, the net benefit is exemplified in *BRCA1/2* mutation carriers (for avoiding unnecessary CPM) and non-carriers (performing necessary CPM). For PredictCBC versions 1B and 2.0B, at the same probability threshold, the net benefit is exemplified in patients with family history (for avoiding unnecessary CPM) and patients without family history (performing necessary CPM).

PredictCBC-1A and PredictCBC-2.0A							
		BRCA1/2 mutation carriers			Non-carriers		
Probability threshold p_t (%)	Unnecessary CPMs needed to detect one necessary CPM*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients using PredictCBC-1A	Additional avoided unnecessary CPMs per 1000 patients using PredictCBC-2.0A	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients using PredictCBC-1A	Additional performed necessary CPMs per 1000 patients using PredictCBC-2.0A
4	24	0.1	0.3	1.9	4.8	115.7	15.3
6	15.7	No benefit	0.0	20.0	0.6	9.3	22.9
8	11.5	3.5	40.6	52.0	No benefit	0.0	9.0
10	9.0	8.5	76.9	50.2	No benefit	0.0	3.4
12	7.3	22.4	164.0	15.0	No benefit	0.0	1.1
PredictCBC-1B and PredictCBC-2.0B							
		Family history			No family history		
Probability threshold p_t (%)	Unnecessary CPMs needed to detect one necessary CPM*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients using PredictCBC-1B	Additional avoided unnecessary CPMs per 1000 patients using PredictCBC-2.0B	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients using PredictCBC-1B	Additional performed necessary CPMs per 1000 patients using PredictCBC-2.0B
4	24	3.4	80.8	5.9	5.4	130.4	0.0
5	19	9.4	177.9	0.0	2.4	46.5	0.1
6	15.7	15.9	248.7	4.0	0.5	7.1	7.5

CPM: contralateral preventive mastectomy;
* The number of unnecessary contralateral mastectomies needed to detect one necessary CPM is calculated by: (1-pt)/pt

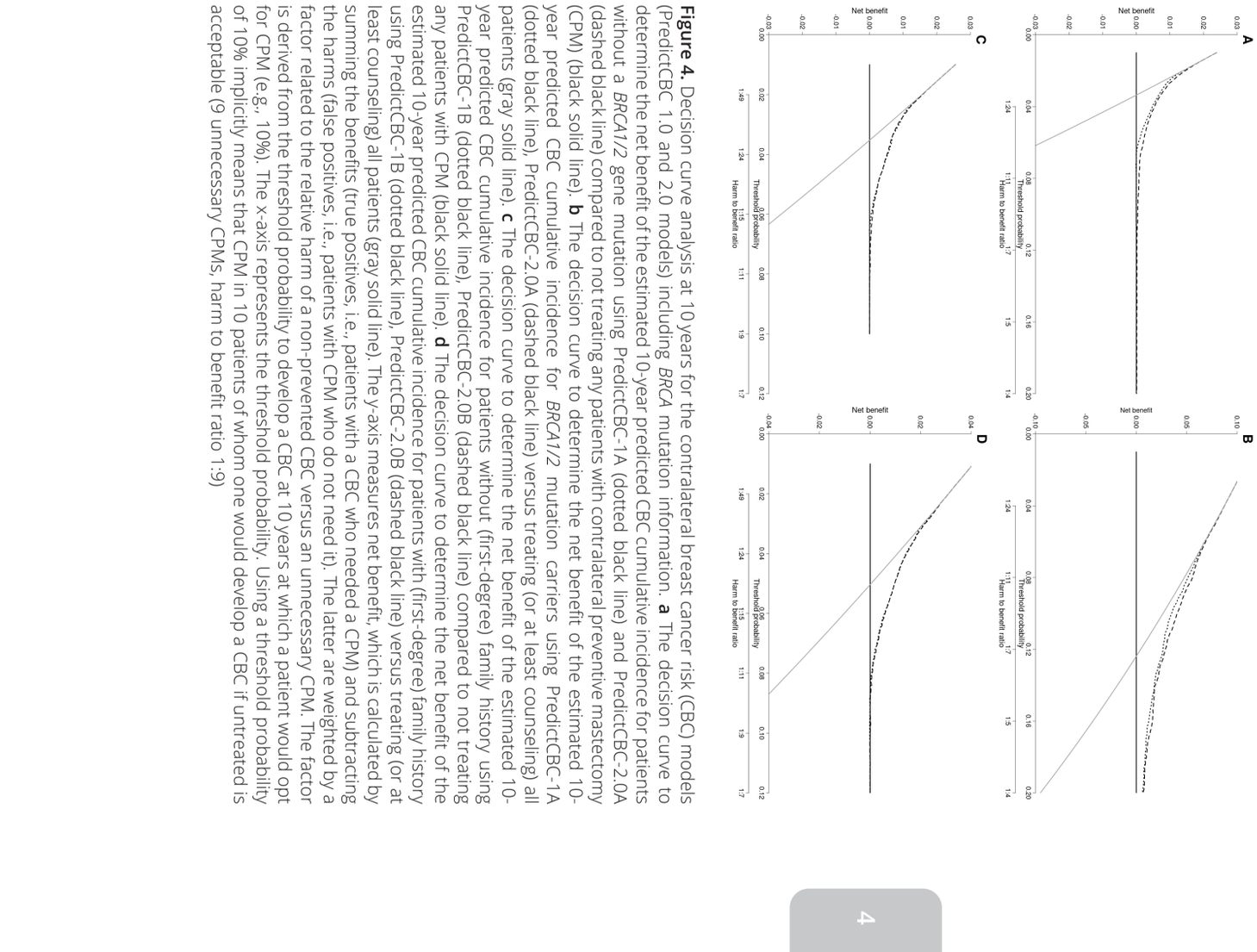


Figure 4. Decision curve analysis at 10 years for the contralateral breast cancer risk (CBC) models (PredictCBC 1.0 and 2.0 models) including *BRCA* mutation information. **a** The decision curve to determine the net benefit of the estimated 10-year predicted CBC cumulative incidence for patients without a *BRCA1/2* gene mutation using PredictCBC-1A (dotted black line) and PredictCBC-2.0A (dashed black line) compared to not treating any patients with contralateral preventive mastectomy (CPM) (black solid line). **b** The decision curve to determine the net benefit of the estimated 10-year predicted CBC cumulative incidence for *BRCA1/2* mutation carriers using PredictCBC-1A (dotted black line), PredictCBC-2.0A (dashed black line) versus treating (or at least counseling) all patients (gray solid line). **c** The decision curve to determine the net benefit of the estimated 10-year predicted CBC cumulative incidence for patients without (first-degree) family history using PredictCBC-1B (dotted black line), PredictCBC-2.0B (dashed black line) compared to not treating any patients with CPM (black solid line). **d** The decision curve to determine the net benefit of the estimated 10-year predicted CBC cumulative incidence for patients with (first-degree) family history using PredictCBC-1B (dotted black line), PredictCBC-2.0B (dashed black line) versus treating (or at least counseling) all patients (gray solid line). The y-axis measures net benefit, which is calculated by summing the benefits (true positives, i.e., patients with CPM who do not need it). The latter are weighted by a factor related to the relative harm of a non-prevented CBC versus an unnecessary CPM. The factor is derived from the threshold probability to develop a CBC at 10 years at which a patient would opt for CPM (e.g., 10%). The x-axis represents the threshold probability. Using a threshold probability of 10% implicitly means that CPM in 10 patients of whom one would develop a CBC if untreated is acceptable (9 unnecessary CPMs, harm to benefit ratio 1:9)

DISCUSSION

We evaluated the potential improvement of CBC risk prediction by adding established genetic (*CHEK2* c.1100delC and PRS-313) and life-style (BMI and parity) factors to the previous PredictCBC models, and used additional follow-up information and new studies to provide more reliable estimates.

The current clinical recommendations of CPM are mostly based on the presence of a pathogenic mutation in *BRCA1/2*[49, 50]. This seems a reasonable approach according to CBC risk predictions based on the PredictCBC models: few non-*BRCA1/2* carriers exceed a 10% 10-year risk threshold. However, approximately 40% of *BRCA1/2* mutation carriers do not reach this threshold either, suggesting that a significant proportion of *BRCA1/2* carriers might be spared CPM. Additional genetic information beyond *BRCA1/2* germline mutation such as the presence of the *CHEK2* c.1110delC variant and PRS-313 might improve decision making.

Currently available CBC models, such as CBCrisk and the Manchester formula, show only moderate discrimination[51]. In addition, the Manchester formula has been shown to systematically overestimate CBC risk[51]. The BOADICEA model, a well-known risk prediction tool to estimate risk of developing first primary BC, also allows the calculation of CBC risk[52-55]. Although BOADICEA includes rare pathogenic variants in moderate and high risk BC susceptibility genes (i.e., *BRCA1*, *BRCA2*, *PALB2*, *ATM* and *CHEK2*, *BARD1*, *RAD51C*, *RAD51D*), and PRS-313, it does not incorporate information on systemic treatment of the primary BC, which are important predictors of CBC risk[56].

A model for prediction of recurrence, the INFLUENCE nomogram, was developed to estimate five-year recurrence risk as well as conditional annual risks of developing a local or regional recurrence based on first BC and treatment characteristics[57]. A more recent version (INFLUENCE 2.0) also provides 5-year individualized predictions for secondary primary breast cancer based on cases older than 50 years at first cancer diagnosis from the NCR nationwide cohort irrespective of their genetic status or testing status using random survival forests[58]. The model provided moderate discrimination (AUC at 5 years: 0.67; 95%CI:0.65–0.68) using internal validation. In our comparable population- and hospital-based Dutch series, EMC and NCR, the AUCs at 5 years of PredictCBC-1A were 0.69 (95%CI:0.64–0.73) and 0.66 (95%CI:0.65–0.67), and of PredictCBC-2.0A 0.71 (95%CI:0.66–0.75) and 0.68 (95%CI:0.66–0.69), respectively. Moreover, INFLUENCE 2.0 is only relevant for the general population, while PredictCBC can also be used in the clinical genetic setting. Notably, we demonstrated that decision making about preventive strategies in clinical practice is unlikely to improve without genetic information.

Our work has some limitations: firstly, some women included in the Dutch studies (providing specific information on family history, *BRCA* mutation or CPM) were also present in our selection of the NCR population, as described previously[15]. Privacy and coding issues prevented linkage at the individual patient level, but based on the hospitals from which the studies recruited, and the age and period criteria used, we calculated a maximum potential overlap of 9%. Secondly, important predictors such as family history, *BRCA1/2* and *CHEK2* c.1110delC status, and PRS-313, were only available in a subset of the women, although the multiple imputation approach should lead to consistent estimates[59-61]. Detailed information about family history would have been useful to improve CBC risk prediction, especially among patients with a mutation in *BRCA1/2* or *CHEK2*. Nonetheless, we considerably increased the number of patients with *BRCA1/2* mutation status and family history information compared to our previous publication (40,343 vs 7,704 and 53,399 vs 30,541 patients with available *BRCA* mutation status and family history information, respectively), and added *CHEK2* c.1110delC, which is a founder mutation present in approximately 0.5–1.6% of individuals of Northern and Eastern European descent and explains the large majority of carriers of *CHEK2* protein truncating variants in these populations[19, 62]. Further validation will be required to investigate how well PredictCBC models predict risk in other populations. In particular, the model was developed in patients of European ancestry and further evaluation and adaptation will be needed to extend PredictCBC models to non-European populations[63, 64]. Future research might also include comparisons of machine learning (ML) methods with classical statistical regression models[65, 66].

The prediction models may be further improved by including additional risk factors. In particular, rare mutations in other breast cancer susceptibility genes, such as *ATM* and *PALB2* are also likely to be associated with an increased risk of CBC[22, 67, 68]. The discrimination provided by the PRS will also improve as more SNPs are added[69, 70]. Prediction performance might also be improved by adding breast density and other risk factors, modelled dynamically in a time dependent fashion[71]. Finally, we wish to emphasize that adequate presentation (e.g., with online tools) of the risk estimates is crucial for effective communication about CBC risk during doctor-patient consultations[72, 73].

CONCLUSIONS

In conclusion, we present an updated version of a previously proposed contralateral breast cancer risk model (PredictCBC) including additional information on breast cancer genetic variants beyond *BRCA1/2*, lifestyle and reproductive factors. PredictCBC-2.0, available online, is based on longer follow-up from a wide range of new European-descent population and hospital-based studies, with satisfactory calibration. PredictCBC

2.0 may be used to tailor clinical decision making towards CPM or alternative preventive strategies, especially when genetic information is available.

Abbreviations

AUC: Area-under-the-ROC-curve; **BC:** Breast cancer; **BCAC:** Breast Cancer Association Consortium;
BMI: Body mass index; **CBC:** Contralateral breast cancer; **CI:** Confidence interval; **CPM:** Contralateral preventive mastectomy; **DCA:** Decision curve analysis; **ER:** Estrogen receptor; **HER2:** Human epidermal growth receptor 2; **ICPW:** Inverse censoring probability weighting; **MICE:** Multiple imputation by chained equations; **PI:** Prediction interval; **PR:** Progesterone receptor; **SEER:** Surveillance, Epidemiology and End Results; **TNM:**TNM Classification of Malignant Tumors.

Acknowledgements

We thank all individuals who took part in these studies and all researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out.

ABCFS thank Maggie Angelakos, Judi Maskiell, Gillian Dite. ABCS thanks the Blood bank Sanquin, The Netherlands. ABCTB Investigators: Christine Clarke, Deborah Marsh, Rodney Scott, Robert Baxter, Desmond Yip, Jane Carpenter, Alison Davis, Nirmala Pathmanathan, Peter Simpson, J. Dinny Graham, Mythily Sachchithananthan. ABCS and BOSOM thank all the collaborating hospitals and pathology departments and many individuals that made this study possible; specifically, we wish to acknowledge: Annegien Broeks, Sten Cornelissen, Frans Hogervorst, Laura van 't Veer, Emiel Rutgers. EMC thanks J.C. Blom-Leenheer, P.J. Bos,C.M.G. Crepin and M. van Vliet for data management. CGPS thanks staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. The Danish Cancer Biobank is acknowledged for providing infrastructure for the collection of blood samples for the cases. HEBCS thanks Johanna Kiiski, Taru A. Muranen, Kristiina Aittomäki, Kirsimari Aaltonen, Karl von Smitten, Irja Erkkilä. The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON) consists of the following Collaborating Centers: Netherlands Cancer Institute (coordinating center), Amsterdam, NL: M.A. Rookus, F.B.L. Hogervorst, M.A. Adank, D.J. Stommel-Jenner, R. de Groot; Erasmus Medical Center, Rotterdam, NL: J.M. Collée, A.M.W. van den Ouweland, M.J. Hooning, I.A. Boere; Leiden University Medical Center, NL: C.J. van Asperen, P. Devilee, R.B. van der Luijt, T.C.T.E.F. van Cronenburg; Radboud University Nijmegen Medical Center, NL: M.R. Wevers, A.R. Mensenkamp; University Medical Center Utrecht, NL: M.G.E.M. Ausems, M.J. Koudijs; Amsterdam UMC, Univ of Amsterdam, NL: I. van de Beek; Amsterdam UMC, Vrije Universiteit Amsterdam, NL: J.J.P. Gille; Maastricht University Medical Center, NL: E.B. Gómez García, M.J. Blok, M. de Boer; University of Groningen, NL: L.P.V. Berger, M.J.E. Mourits, G.H. de Bock; The Netherlands Comprehensive Cancer Organisation (IKNL): J. Verloop; The nationwide network and registry of histo- and cytopathology in The Netherlands (PALGA): E.C. van den Broek. HEBON thanks the study participants and the registration teams of IKNL and PALGA for part of the data collection. KARMA thanks the Swedish Medical Research Counsel. LMBC thanks Gilian Peuteman, Thomas Van Brussel, EvyVanderheyden and Kathleen Corthouts. MARIE thanks Petra Seibold, Nadia Obi, Sabine Behrens, Ursula Eilber and Muhabbet Celik ORIGO thanks E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The authors thank the registration team of the Netherlands Comprehensive Cancer Organisation (IKNL) for the collection of data for the Netherlands Cancer Registry as well as IKNL staff for scientific advice.

PBCS thanks Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner. The ethical approval for the POSH study is MREC /00/6/69, UKCRN ID: 1137. We thank the SEARCH and EPIC teams. SKKDKFZS thanks all study participants, clinicians, family doctors, researchers and technicians for their contributions and commitment to this study. SZBCS thanks Ewa Putresza. UBCS thanks all study participants, the ascertainment, laboratory and research informatics teams at Huntsman Cancer Institute and Intermountain Healthcare, and Justin Williams, Brandt Jones, Myke Madsen, Melissa Cessna, Stacey Knight and Kerry Rowe for their important contributions to this study. Special thanks to Stefano Bottelli for his R programming support.

Availability of data and materials

All data relevant to this report are included in this published article and its supplementary information files. The datasets analyzed during the current study are not publicly available due to protection of participant privacy and confidentiality. Pseudomised data sets that were used in the analyses can be requested from the Netherlands Cancer Registry, the Netherlands Cancer Institute, ErasmusMC, and the Breast Cancer Assocation Consortium.

Funding

This work is supported by the Alpe d'HuZes/Dutch Cancer Society (KWF Kankerbestrijding) project 6253.

BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation

Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report. Additional funding for BCAC is provided via the Confluence project which is funded with intramural funds from the National Cancer Institute Intramural Research Program, National Institutes of Health.

The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. BOSOM was supported by the Dutch Cancer Society grant numbers DCS-NKI 2001-2423, DCS-NKI 2007-3839, and DCSNKI 2009-4363; the Cancer Genomics Initiative; and notary office Spier & Hazenberg for the coding procedure. The BREast Oncology GAlician Network (BREOGAN) is funded by Acción Estratégica de Salud del Instituto de Salud Carlos III FIS PI12/02125/Cofinanciado and FEDER PI17/00918/Cofinanciado FEDER; Acción Estratégica de Salud del Instituto de Salud Carlos III FIS Intrasalud (PI13/01136); Programa Grupos Emergentes, Cancer Genetics Unit, Instituto de Investigacion Biomedica Galicia Sur. Xerencia de Xestion Integrada de Vigo-SERGAS, Instituto de Salud Carlos III, Spain; Grant 10CSA012E, Consellería de Industria Programa Sectorial de Investigación Aplicada, PEME I + D e I + D Suma del Plan Gallego de Investigación, Desarrollo e Innovación Tecnológica de la Consellería de Industria de la Xunta de Galicia, Spain; Grant EC11-192. Fomento de la Investigación Clínica Independiente, Ministerio de Sanidad, Servicios Sociales e Igualdad, Spain; and Grant FEDER-Innterconecta. Ministerio de Economia y Competitividad, Xunta de Galicia, Spain. The EMC was supported by grants from Alpe d'HuZes/Dutch Cancer Society NKI2013-6253 and from Pink Ribbon 2012.WO39.C143. The HEBCS was financially supported by the Helsinki University Hospital Research Fund, the Finnish Cancer Society, and the Sigrid Juselius Foundation. The HEBON study is supported by the Dutch Cancer Society grants NKI1998-1854, NKI2004-3088, NKI2007-3756, NKI 12535, the Netherlands Organisation of Scientific Research grant NWO 91109024, the Pink Ribbon grants 110005 and 2014-187.WO76, the BBMRI grant NWO 184.021.007/CP46, and the Transcan grant JTC 2012 Cancer 12-054.

Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantzows foundation. The KARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. LMBC is supported by the 'Stichting tegen Kanker'. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419, 110826, 110828], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MEC was support by NIH grants CA63464, CA54281, CA098758, CA132839 and CA164973. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The Netherlands Cancer Registry is hosted by the Netherlands Comprehensive Cancer Organisation (IKNL) and financed by the Dutch Ministry of Health, Welfare and Sports. The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. The POSH study is funded by Cancer Research UK (grants C1275/A11699, C1275/C22524, C1275/A19187, C1275/A15956 and Breast Cancer Campaign 2010PR62, 2013PR044). SKKDKFZS is supported by the DKFZ. The SZBCS was supported by Grant PBZ_KBN_122/P05/2004 and the program of the Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in 2019-2022 project number 002/RID/2018/19 amount of financing 12 000 000 PLN. UBCS was supported by funding from National Cancer Institute (NCI) grant R01 CA163353 (to N.J. Camp) and the Women's Cancer Center at the Huntsman Cancer Institute (HCI). Data collection for UBCS was supported by the Utah Population Database, Intermountain Healthcare and the Utah Cancer Registry which is funded by the NCI's SEER Program (HHSN261201800016I), the US Centers for Disease Control and Prevention's National Program of Cancer Registries (NU58DP006320), with additional support from the University of Utah and Huntsman Cancer Foundation.

Authors' contributions

MKS, MJH conceived the study in collaboration with EWS and MH. DG performed the statistical analysis. DG, MKS, MJH, EWS and MH interpreted the results and drafted the manuscript. All remaining authors contributed to critical revision and editing of the final version of the manuscript for publication. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Each study was approved by its institutional ethical review board.

Data availability statement

The datasets analyzed during the current study are not publicly available due to protection of participant privacy and confidentiality, and ownership of the contributing institutions, but may be made available in an anonymized form via the corresponding author on reasonable request and after approval of the involved institutions

Competing interests

The authors declare that they have no competing interests.

REFERENCES

1. Chen Y, Thompson W, Semenciw R, *et al.* Epidemiology of contralateral breast cancer. Cancer Epidemiol Biomarkers Prev 1999;8(10):855-61.
2. Gao X, Fisher SG, Emami B. Risk of second primary cancer in the contralateral breast in women treated for early-stage breast cancer: a population-based study. Int J Radiat Oncol Biol Phys 2003;56(4):1038-45.
3. Curtis RE, Ron E, Hankey BF, *et al.* New Malignancies Following Breast Cancer. In. *New malignancies among cancer survivors: SEER Cancer Registries, 1973-2000*, 181-205.
4. Yu GP, Schantz SP, Neugut AI, *et al.* Incidences and trends of second cancers in female breast cancer patients: a fixed inception cohort-based analysis (United States). Cancer Causes Control 2006;17(4):411-20.
5. Soerjomataram I, Louwman WJ, Lemmens VE, *et al.* Risks of second primary breast and urogenital cancer following female breast cancer in the south of The Netherlands, 1972-2001. Eur J Cancer 2005;41(15):2331-7.
6. Schaapveld M, Visser O, Louwman WJ, *et al.* The impact of adjuvant therapy on contralateral breast cancer risk and the prognostic significance of contralateral breast cancer: a population based study in the Netherlands. Breast Cancer Res Treat 2008;110(1):189-97.
7. Tuttle TM, Habermann EB, Grund EH, *et al.* Increasing use of contralateral prophylactic mastectomy for breast cancer patients: a trend toward more aggressive surgical treatment. J Clin Oncol 2007;25(33):5203-9.
8. Narod SA. Bilateral breast cancers. Nat Rev Clin Oncol 2014;11(3):157-66.
9. Metcalfe K, Gershman S, Ghadirian P, *et al.* Contralateral mastectomy and survival after breast cancer in carriers of BRCA1 and BRCA2 mutations: retrospective analysis. BMJ 2014;348:g226.
10. Xiong Z, Yang L, Deng G, *et al.* Patterns of Occurrence and Outcomes of Contralateral Breast Cancer: Analysis of SEER Data. J Clin Med 2018;7(6).
11. Wong SM, Freedman RA, Sagara Y, *et al.* Growing Use of Contralateral Prophylactic Mastectomy Despite no Improvement in Long-term Survival for Invasive Breast Cancer. Ann Surg 2017;265(3):581-589.
12. Murphy JA, Milner TD, O'Donoghue JM. Contralateral risk-reducing mastectomy in sporadic breast cancer. Lancet Oncol 2013;14(7):e262-9.
13. Basu NN, Hodson J, Chatterjee S, *et al.* The Angelina Jolie effect: Contralateral risk-reducing mastectomy trends in patients at increased risk of breast cancer. Sci Rep 2021;11(1):2847.
14. Domchek SM. Risk-Reducing Mastectomy in BRCA1 and BRCA2 Mutation Carriers: A Complex Discussion. JAMA 2019;321(1):27.
15. Giardiello D, Steyerberg EW, Hauptmann M, *et al.* Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res 2019;21(1):144.
16. Basu NN, Ross GL, Evans DG, *et al.* The Manchester guidelines for contralateral risk-reducing mastectomy. World J Surg Oncol 2015;13:237.
17. Chowdhury M, Euhus D, Onega T, *et al.* A model for individualized risk prediction of contralateral breast cancer. Breast Cancer Res Treat 2017;161(1):153-160.
18. Chowdhury M, Euhus D, Arun B, *et al.* Validation of a personalized risk prediction model for contralateral breast cancer. Breast Cancer Res Treat 2018;170(2):415-423.
19. Weischer M, Nordestgaard BG, Pharoah P, *et al.* CHEK2*1100delC heterozygosity in women with breast cancer associated with early death, breast cancer-specific death, and increased risk of a second breast cancer. J Clin Oncol 2012;30(35):4308-16.
20. Akdeniz D, Schmidt MK, Seynaeve CM, *et al.* Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. Breast 2019;44:1-14.
21. Robson ME, Reiner AS, Brooks JD, *et al.* Association of Common Genetic Variants With Contralateral Breast Cancer Risk in the WECARE Study. J Natl Cancer Inst 2017;109(10).
22. Fanale D, Incorvaia L, Filorizzo C, *et al.* Detection of Germline Mutations in a Cohort of 139 Patients with Bilateral Breast Cancer by Multi-Gene Panel Testing: Impact of Pathogenic Variants in Other Genes beyond BRCA1/2. Cancers (Basel) 2020;12(9).
23. Kramer I, Hooning MJ, Mavaddat N, *et al.* Breast Cancer Polygenic Risk Score and Contralateral Breast Cancer Risk. Am J Hum Genet 2020;107(5):837-848.
24. Lakeman IMM, van den Broek AJ, Vos JAM, *et al.* The predictive ability of the 313 variant-based polygenic risk score for contralateral breast cancer risk prediction in women of European ancestry with a heterozygous BRCA1 or BRCA2 pathogenic variant. Genet Med 2021; 10.1038/s41436-021-01198-7.
25. Mavaddat N, Michailidou K, Dennis J, *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019;104(1):21-34.
26. Akdeniz D, Klaver MM, Smith CZA, *et al.* The impact of lifestyle and reproductive factors on the risk of a second new primary cancer in the contralateral breast: a systematic review and meta-analysis. Cancer Causes Control 2020;31(5):403-416.
27. Pijpe A, Manders P, Brohet RM, *et al.* Physical activity and the risk of breast cancer in BRCA1/2 mutation carriers. Breast Cancer Res Treat 2010;120(1):235-44.
28. Riegman PH, van Veen EB. Biobanking residual tissues. Hum Genet 2011;130(3):357-68.
29. Foundation Federation of Dutch Medical Scientific Societies. Human Tissue and Medical Research: Code of Conduct for responsible use. 2011, https://www.federa.org/sites/default/files/images/print_version_code_of_conduct_english.pdf.
30. van den Broek AJ, Schmidt MK, van 't Veer LJ, *et al.* Prognostic Impact of Breast-Conserving Therapy Versus Mastectomy of BRCA1/2 Mutation Carriers Compared With Noncarriers in a Consecutive Series of Young Breast Cancer Patients. Ann Surg 2019;270(2):364-372.
31. Buuren Sv. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press; 2012.
32. Resche-Rigon M, White IR, Bartlett JW, *et al.* Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Stat Med 2013;32(28):4890-905.
33. Van Buuren S. *Flexible imputation of missing data*. Second ed: Chapman and Hall/CRC; 2018.
34. Geskus RB. Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. Biometrics 2011;67(1):39-49.
35. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics 1983;39(2):499-503.
36. Schmidt MK, Tollenaar RA, de Kemp SR, *et al.* Breast cancer survival and tumor characteristics in premenopausal women carrying the CHEK2*1100delC germline mutation. J Clin Oncol 2007;25(1):64-9.
37. Schmidt MK, Hogervorst F, van Hien R, *et al.* Age- and Tumor Subtype-Specific Breast Cancer Risk Estimates for CHEK2*1100delC Carriers. J Clin Oncol 2016;34(23):2750-60.
38. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external

validation. J Clin Epidemiol 2016;69:245-7.
39. Austin PC, van Klaveren D, Vergouwe Y, *et al.* Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol 2016;79:76-85.
40. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016;35(2):214-26.
41. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 2013;32(30):5381-97.
42. Brentnall AR, Cuzick J. Risk Models for Breast Cancer and Their Validation. Stat Sci 2020;35(1):14-30.
43. Austin PC, Putter H, Giardiello D, *et al.* Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagn Progn Res 2022;6(1):2.
44. Collins GS, Reitsma JB, Altman DG, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015;162(10):735-6.
45. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74.
46. Kerr KF, Brown MD, Zhu K, *et al.* Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. J Clin Oncol 2016;34(21):2534-40.
47. Vickers AJ, Cronin AM, Elkin EB, *et al.* Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008;8:53.
48. Heemskerk-Gerritsen BA, Rookus MA, Aalfs CM, *et al.* Improved overall survival after contralateral risk-reducing mastectomy in BRCA1/2 mutation carriers with a history of unilateral breast cancer: a prospective analysis. Int J Cancer 2015;136(3):668-77.
49. Balmana J, Diez O, Rubio IT, *et al.* BRCA in breast cancer: ESMO Clinical Practice Guidelines. Ann Oncol 2011;22 Suppl 6:vi31-4.
50. Rutgers EJT. Is prophylactic mastectomy justified in women without BRCA mutation? Breast 2019;48 Suppl 1:S62-S64.
51. Giardiello D, Hauptmann M, Steyerberg EW, *et al.* Prediction of contralateral breast cancer: external validation of risk calculators in 20 international cohorts. Breast Cancer Res Treat 2020;181(2):423-434.
52. Antoniou AC, Pharoah PP, Smith P, *et al.* The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer 2004;91(8):1580-90.
53. Antoniou AC, Cunningham AP, Peto J, *et al.* The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Br J Cancer 2008;98(8):1457-66.
54. Lee AJ, Cunningham AP, Tischkowitz M, *et al.* Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. Genet Med 2016;18(12):1190-1198.
55. Carver T, Hartley S, Lee A, *et al.* CanRisk Tool-A Web Interface for the Prediction of Breast and Ovarian Cancer Risk and the Likelihood of Carrying Genetic Pathogenic Variants. Cancer Epidemiol Biomarkers Prev 2021;30(3):469-473.
56. Kramer I, Schaapveld M, Oldenburg HSA, *et al.* The Influence of Adjuvant Systemic Regimens on Contralateral Breast Cancer Risk and Receptor Subtype. J Natl Cancer Inst 2019;111(7):709-718.
57. Witteveen A, Vliegen IM, Sonke GS, *et al.* Personalisation of breast cancer follow-up: a time-dependent prognostic nomogram for the estimation of annual risk of locoregional recurrence in early breast cancer patients. Breast Cancer Res Treat 2015;152(3):627-36.
58. Volkel V, Hueting TA, Draeger T, *et al.* Improved risk estimation of locoregional recurrence, secondary contralateral tumors and distant metastases in early breast cancer: the INFLUENCE 2.0 model. Breast Cancer Res Treat 2021; 10.1007/s10549-021-06335-z.
59. Nieboer D, Vergouwe Y, Ankerst DP, *et al.* Improving prediction models with new markers: a comparison of updating strategies. BMC Med Res Methodol 2016;16(1):128.
60. Madley-Dowd P, Hughes R, Tilling K, *et al.* The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol 2019;110:63-73.
61. Collins GS, Altman DG. Predicting the 10 year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. BMJ 2012;344:e4181.
62. Breast Cancer Association C, Dorling L, Carvalho S, *et al.* Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. N Engl J Med 2021;384(5):428-439.
63. Ho WK, Tan MM, Mavaddat N, *et al.* European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. Nat Commun 2020;11(1):3833.
64. Evans DG, van Veen EM, Byers H, *et al.* The importance of ethnicity: Are breast cancer polygenic risk scores ready for women who are not of White European origin? Int J Cancer 2021; 10.1002/ijc.33782.
65. Christodoulou E, Ma J, Collins GS, *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.
66. Giardiello D, Antoniou AC, Mariani L, *et al.* Letter to the editor: a response to Ming's study on machine learning techniques for personalized breast cancer risk prediction. Breast Cancer Res 2020;22(1):17.
67. Thompson D, Easton D. The genetic epidemiology of breast cancer genes. J Mammary Gland Biol Neoplasia 2004;9(3):221-36.
68. Reiner AS, Sisti J, John EM, *et al.* Breast Cancer Family History and Contralateral Breast Cancer Risk in Young Women: An Update From the Women's Environmental Cancer and Radiation Epidemiology Study. J Clin Oncol 2018;36(15):1513-1520.
69. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet 2018;19(9):581-590.
70. Wald NJ, Old R. The illusion of polygenic disease risk prediction. Genet Med 2019; 10.1038/s41436-018-0418-5.
71. Knight JA, Blackmore KM, Fan J, *et al.* The association of mammographic density with risk of contralateral breast cancer and change in density with treatment in the WECARE study. Breast Cancer Res 2018;20(1):23.
72. Van Belle V, Van Calster B. Visualizing Risk Prediction Models. PLoS One 2015;10(7):e0132614.
73. Bonnett LJ, Snell KIE, Collins GS, *et al.* Guide to presenting clinical prediction models for use in clinical settings. BMJ 2019;365:l737.

SUPPLEMENTARY MATERIALS

1. Data and patient selection

For this study we used data from six main sources available from national and international collaborations including nationwide registry data, as well as hospital-based studies with more detailed information on relevant prediction factors[1-5]. Briefly, the six main sources were: (1) The Breast Cancer Association Consortium (BCAC), which is an international consortium of 106 studies comprising 186,594 patients (data version August 2019) with a primary breast cancer (BC) diagnosed between 1939 and 2018[1]. In our previous study, 16 studies were selected to develop PredictCBC models. In this study, two studies were additionally included in the dataset to develop PredictCBC-2.0 models[6]; (2) The Amsterdam Breast Cancer Study (ABCS) containing 2,763 patients diagnosed with a first BC at the Netherlands Cancer Institute – Antoni van Leeuwenhoek (NKI-AVL) hospital in Amsterdam from 2003 to 2013[2]; (3) The Breast Cancer Outcome Study of Mutation carriers (BOSOM), which is a Dutch consecutive series of 7,105 patients with invasive BC treated for their primary BC in ten centers throughout the Netherlands between 1970 and 2003; in this study 94% of patients were genotyped for *BRCA1/2* germline mutations[3]; (4) The Erasmus Medical Center (EMC) study including patients diagnosed with BC between 1989 and 2013 who were treated at the EMC in Rotterdam; for this study, complete follow-up was obtained for 3,483 patients who had been diagnosed between 2000 and 2009;(5) The Netherlands Cancer Registry (NCR), which is an ongoing nationwide population-based data registry of all newly diagnosed cancer patients in the Netherlands since 1989[4]. We included patients diagnosed between 2003 and 2015, a period for which sufficient follow-up and receptor status information were available[4, 5];(6) Hereditary Breast and Ovarian cancer study, the Netherlands (HEBON) study is an ongoing nationwide Dutch study among members of *BRCA1/2* families in the Netherlands, including 16,617 BC patients diagnosed between 1953 and 2017[7]. The general design includes a retrospective cohort because the *BRCA1/2* DNA test was available from 1995, with a prospective follow-up. *BRCA1/2* families were identified through ten centers (nine Clinical Genetic Centers/Family Cancer Clinics and the Foundation for the Detection of Hereditary Tumors). The eligibility criteria applied in each data source is reported in **Table S1**. Data were harmonized by recoding each of the main datasets by the responsible data managers according to a standardized data dictionary. We performed checks for data consistency and validity centrally.

We extracted the following information: *BRCA1/2* germline mutation, family history (first degree) of primary BC, *CHEK2* c.1100delC, polygenic risk score (PRS) (derived from BCAC), body mass index (BMI), parity and regarding primary BC diagnosis: age, nodal status, size, grade, morphology, estrogen-receptor (ER) status, progesterone-receptor (PR), human epidermal growth factor receptor 2 (HER2) status, administration of adjuvant or neoadjuvant chemotherapy, adjuvant endocrine therapy, adjuvant trastuzumab therapy, radiotherapy[2, 8, 9]. We excluded PR status and TNM stage of the primary BC due to collinearity with ER status and the size of the primary tumor, respectively. In current clinical practice, only patients with ER-positive tumors receive endocrine therapy and only patients with HER2-positive tumors receive trastuzumab; these co-occurrences were considered in the model by using composite categorical variables. A description of the studies included in the analyses is provided in **Supplementary Table 2**. Follow-up started three months after invasive first primary BC diagnosis, to exclude synchronous contralateral breast cancer (CBC), and ended at date of CBC, distant metastasis (but not loco-regional relapse), CPM, or last date of follow-up (due to death, being lost to follow-up, or end of study), whichever occurred first. We considered that after loco-regional relapse, a woman would be still at risk for CBC as treatment for loco-regional relapse would not affect CBC unless adjuvant systemic treatment was given. Distant metastasis was considered as a competing risk because most of the patients receive systemic therapies after developing distant metastasis.

Age at first primary BC seemed to have a non-linear relation with CBC. Using splines, we observed that CBC risk increased with age till around 60 years old and declined afterwards. Therefore, we used a linear spline with a knot at 60 years in the prediction model. The use of this linear spline was a good compromise to address the non-linear relation between CBC risk and age across the different baseline risks in all the studies, with different age distributions and selections (one study included only women aged under 50 years). Moreover, the observed non-linear relation resembled the shape of age-related BC incidence curves with an increased risk until menopausal age followed by a decrease (Clemmensen's hook)[10].

2. Multiple imputation of missing values

The percentage of missing values across the predictors varied between 3.2% and 84% for morphology of first primary BC and *BRCA* mutation, respectively. In the individual patient data (IPD), both sporadic and systematic missing may occur. The former are missing values within a study, the latter are values missing for all individuals within a particular study[11-13].

For our analyses, we used five imputed datasets based on the multiple imputation chained equations (MICE) using 50 iterations. The visit sequence of the variables was in ascending order of the number of missing values. This technique improves the accuracy and the statistical power assuming missing is at random (MAR). In the imputation procedure, we also used the year of first primary BC diagnosis since this information provides a better correlation structure among covariates used as predictors in the imputation model. Since there were systematic missing data, we used the imputation model based on the stratified multiple imputation strategy (SMI). In this approach,

the variable identifying the study was used as covariate to improve substantially the imputation especially for the systematic missing predictors that might occur in the IPD from multiple studies[13]. Continuous, binary, and multiple categorical variables were imputed using predictive mean matching, binary and multinomial logistic regression, respectively. Time-to-event outcome defined as time to CBC, time to death, and time to distant metastasis were included in the imputation process through the Nelson-Aalen cumulative hazard estimator[14]. For every variable with missing data, every imputation model selects predictors based on correlation structure underlying the data. We recoded the variables chemotherapy and morphology after imputation. Information about neoadjuvant and adjuvant chemotherapy were separately imputed. Then, we created a chemotherapy variable by combining the variables for neoadjuvant and adjuvant chemotherapy in every imputed dataset. Morphology of primary tumor was imputed by keeping all original categories ('Lobular', 'Ductal', 'Mixed (ductal and lobular)' and 'Other'). After multiple imputation, we created two categories 'Lobular including mixed' and 'Ductal including other' to mitigate possible overfitting due to the small numbers of patients with 'Mixed' and 'Other' categories. Since in current clinical practice, only estrogen receptor (ER) positive patients receive endocrine therapy and only human epidermal growth factor receptor 2 (HER2) positive patients receive trastuzumab, composite categorical factors of ER and endocrine therapy and of HER2 and trastuzumab therapy were considered in the model building. However, in our data, 1% of patients with 97 CBC events were coded as ER-negative treated with endocrine therapy and 0.1% of patients with 11 CBC events were coded as HER2-negative treated with trastuzumab therapy. In every imputed dataset, we recoded those patients as ER-positive treated with endocrine treatment and HER2-positive treated with trastuzumab since the largest proportion of patients (67%) were ER-positive treated with endocrine therapy and 60% were HER2-positive treated with trastuzumab in the complete data.

We used the R package mice (version 3.13.0) to impute our data and combine the estimates using Rubin's rules.

3. Formula to estimate the contralateral breast cancer risk using PredictCBC-2.0A

Our developed model is a subdistributional proportional hazard Fine and Gray model. The estimated cumulative incidence of CBC was estimated using the following formula:

$$F(t) = 1 - \left\{[S_0(t)]^{\exp(\boldsymbol{LP})}\right\}$$

Where *t* is the time (in years) since primary BC, $F(t)$ is the cumulative incidence of CBC and $S_0(t)$ is the probability to survive beyond for baseline covariate values. To calculate the predicted CBC cumulative incidence, we used the event-free baseline probability of the Dutch Cancer Registry. The baseline survival estimates according to the model and time *t* are:

$S_0(5) = 0.985$

$S_0(10) = 0.971$

And

Linear Predictor (LP) =

– 0.303 + 0.003× Age – 0.031 × Age' + 0.011 × BMI – 0.0812 × Parity + 0.157 × I[Family history = Yes] + 1.566 × I[*BRCA* = *BRCA1*] + 1.128 × I[*BRCA* = *BRCA2*] + 0.938 × I[*CHEK2* c.1100delC] + 0.398 × PRS-313 – 0.011 × I[Nodal status = positive] – 0.089 × I[Size of PBC = (2,5] *cm*] + 0.201 × I[Size of PBC = greater than 5 *cm*] + 0.186 × I[Morphology of PBC = lobular including mixed] – 0.069 × I[Grade of PBC = moderately differentiated] – 0.163 × I[Grade of PBC = poorly/undifferentiated]– 0.285 × I[Chemotherapy = yes] + 0.065 × I[Radiotherapy to the breast = yes] + 0.428 × I[ER-negative without endocrine therapy] + 0.668 × I[ER-positive without endocrine therapy]+ 0.203 × I[HER2-negative without trastuzumab] + 0.111 × I[HER2-positive without trastuzumab]

Where Age' = max(Age – 60, 0), with age in years

4. Formula to estimate the contralateral breast cancer risk in using PredictCBC-2.0B

The formula for the alternative model is reported below. Baseline survival estimates according to the model and time *t* are:

$S_0(5) = 0.984$

$S_0(10) = 0.970$

And

– 0.160 – 0.002 × Age – 0.029 × Age' + 0.011 × BMI – 0.0728 × Parity + 0.304 × I[Family history = Yes] – 0.013× I[Nodal status = positive] + 0.011 × I[Size of PBC = (2,5] *cm*] + 0.198 × I[Size of PBC = greater than 5 *cm*] + 0.158 × I[Morphology of PBC = lobular including mixed] – 0.017 × I[Grade of PBC = moderately differentiated] – 0.055 × I[Grade of PBC = poorly/undifferentiated]– 0.293 × I[Chemotherapy = yes] – 0.055 × I[Radiotherapy to the breast = yes] + 0.578 × I[ER-negative without endocrine therapy] + 0.661 × I[ER-positive without endocrine therapy]+ 0.262 × I[HER2-negative without trastuzumab] + 0.133 × I[HER2-positive without trastuzumab]

Where Age' = max(Age – 60, 0), with age in years.

Supplementary Table 1: Patient characteristics in the different data sources.

	Source of data					
	ABCS	BCAC‡	BOSOM	EMC	HEBON	NCR
Number of patients	2,763	186,594	7,105	3,483	16,617	160,861
Eligibility criteria, *number of patients excluded*						
Studies from Asian countries	-	7,146	-	-	-	-
Patients of non-European descent	74	51,328	-	-	-	-
Patients younger than 18 years old	-	4	-	-	-	-
Year of PBC diagnosis before 1990	-	4,014	3,126	-	1,132	-
Year of PBC diagnosis missing	-	15,435	-	-	2	-
PBC stage 0	123	38	2	-	-	-
PBC stage IV	149	1,811	104	-	115	7,774
Patients did not undergo surgery	24	1,247	43	5	293	9,278
Number of eligible patients	2,393	105,571	3,830	3,478	15,075	143,809
No follow-up or follow-up less than 3 months	173	15,804	70	88	2,382*	3,396
Familiar breast cancer studies	-	6,739	-	-	-	-
Studies with less than 10 CBC events	-	37,994	-	-	-	-
Number of patients included in the analysis (number of patients with CBC)	2,220 (44)	45,034 (1,001)	3,760 (288)	3,390 (221)	12,693 (918)	140,413 (5,753)
Total number of patients included in the analysis (number of CBC)	207,510 (8,225 of which 6,828 invasive and 1,397 in situ)					

Abbreviations:
ABCS: Amsterdam Breast Cancer Study; BCAC: Breast Cancer Association Consortium. ‡BCAC is composed of 106 studies world-wide. The 45,034 patients selected for the analysis came from 18 studies; BOSOM: Breast Cancer Outcome Study of Mutation carriers; EMC: Erasmus Medical Center; HEBON: Hereditary Breast and Ovarian cancer study Netherlands. *1,433 tested for *BRCA1/2* germline mutation after CBC or preventive mastectomy; NCR: Netherlands Cancer Registry; PBC: primary breast cancer; CBC: contralateral breast cancer

Supplementary Table 2: available online.

Supplementary Table 3: List of BCAC studies (including ABCS source) with the corresponding country and geographic area. For studies in which the number of contralateral breast cancer events was insufficient for external validation, the geographic area was used.

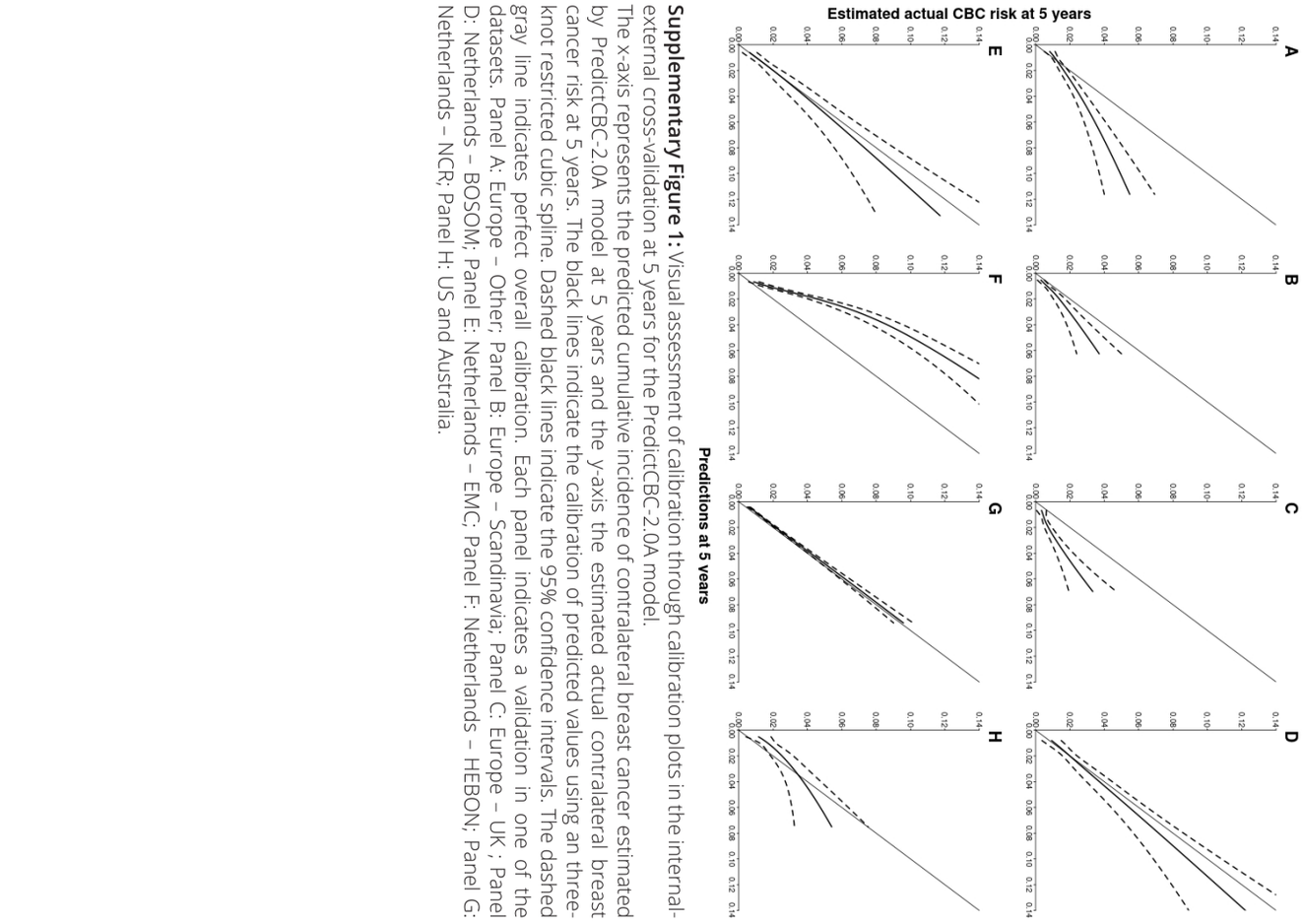
Study	Country	Geographic area or study
ABCS	Netherlands	Europe - Other
ABCFS	Australia	United States and Australia
BBCC	Germany	Europe - Other
BREOGAN	Spain	Europe - Other
CGPS	Denmark	Europe - Scandinavia
HEBCS	Finland	Europe - Scandinavia
KARBAC	Sweden	Europe - Scandinavia
KARMA	Sweden	Europe - Scandinavia
LMBC	Belgium	Europe - Other
MARIE	Germany	Europe - Other
MEC	United States	United States and Australia
ORIGO	Netherlands	Europe - Other
PBCS	Poland	Europe - Other
PKARMA	Sweden	Europe - Scandinavia
POSH	United Kingdom	Europe - United Kingdom
SEARCH	United Kingdom	Europe - United Kingdom
SKKDKFZS	Germany	Europe - Other
SZBCS	Poland	Europe - Other
UBCS	United States	United States and Australia

Supplementary Table 4: available online (Patient and primary breast cancer characteristics per study).

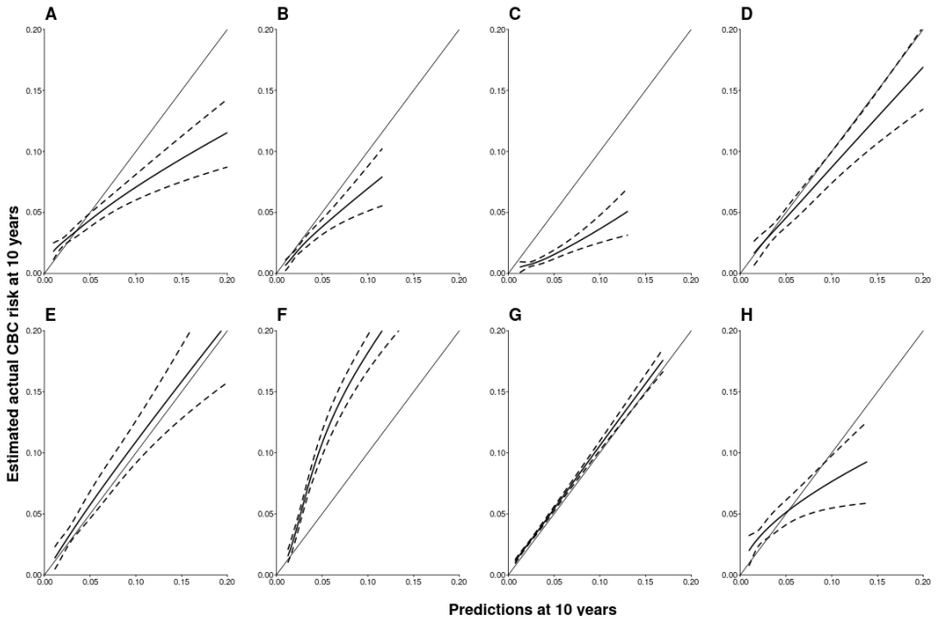
Supplementary Table 5: Clinical utility of the 5-year contralateral breast cancer risk prediction models (PredictCBC-1A with PredictCBC-2.0A and PredictCBC-1B with PredictCBC-2.0B). For PredictCBC versions 1A and 2.0A, at the same probability threshold, the net benefit is exemplified in *BRCA1/2* mutation carriers (for avoiding unnecessary CPM) and non-carriers (performing necessary CPM). For PredictCBC versions 1B and 2.0B, at the same probability threshold, the net benefit is exemplified in patients with family history (for avoiding unnecessary CPM) and patients without family history (performing necessary CPM).

PredictCBC-1A and PredictCBC-2.0A							
		BRCA1/2 mutation carriers			Non-carriers		
Probability threshold p_t (%)	Unnecessary CPMs needed to detect one necessary CPM*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients using PredictCBC-1A	Additional avoided unnecessary CPMs per 1000 patients using PredictCBC-2.0A	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients using PredictCBC-1A	Additional performed necessary CPMs per 1000 patients using PredictCBC-2.0A
3	32.3	0.2	6.0	0.0	0.6	19.7	210.9
4	24.0	1.9	44.4	16.4	No benefit	0.0	129.4
5	19.0	3.4	64.1	66.7	No benefit	0.0	56.9
6	15.7	9.4	146.6	34.1	No benefit	0.0	0.0
PredictCBC-1B and PredictCBC-2.0B							
		Family history			No family history		
Probability threshold p_t (%)	Unnecessary CPMs needed to detect one necessary CPM*	Net benefit versus treat all patients with CPM (per 1000)	Avoided unnecessary CPMs per 1000 patients using PredictCBC-1B	Additional avoided unnecessary CPMs per 1000 patients using PredictCBC-2.0B	Net benefit versus treat none (per 1000)	Performed necessary CPMs per 1000 patients using PredictCBC-1B	Additional performed necessary CPMs per 1000 patients using PredictCBC-2.0B
2	49	2.3	115.1	0.0	3.4	168.1	0.0
2.5	39	5.7	200.4	0.0	1.8	70.1	0.0
3	32.3	3.6	258.3	0.0	0.6	19.9	0.3

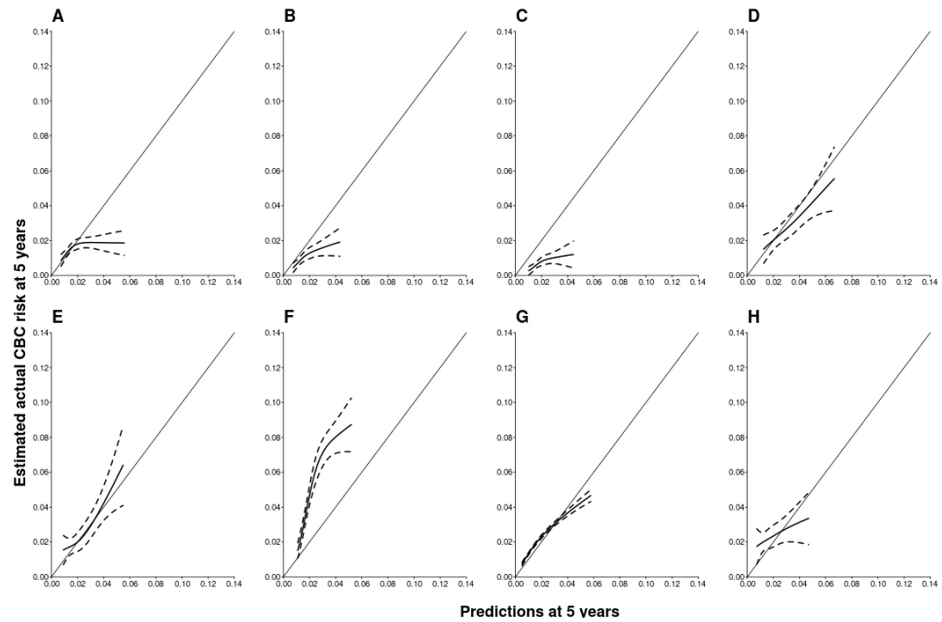
CPM: contralateral preventive mastectomy;
* The number of unnecessary contralateral mastectomies needed to detect one necessary CPM is calculated by: (1-pt)/pt



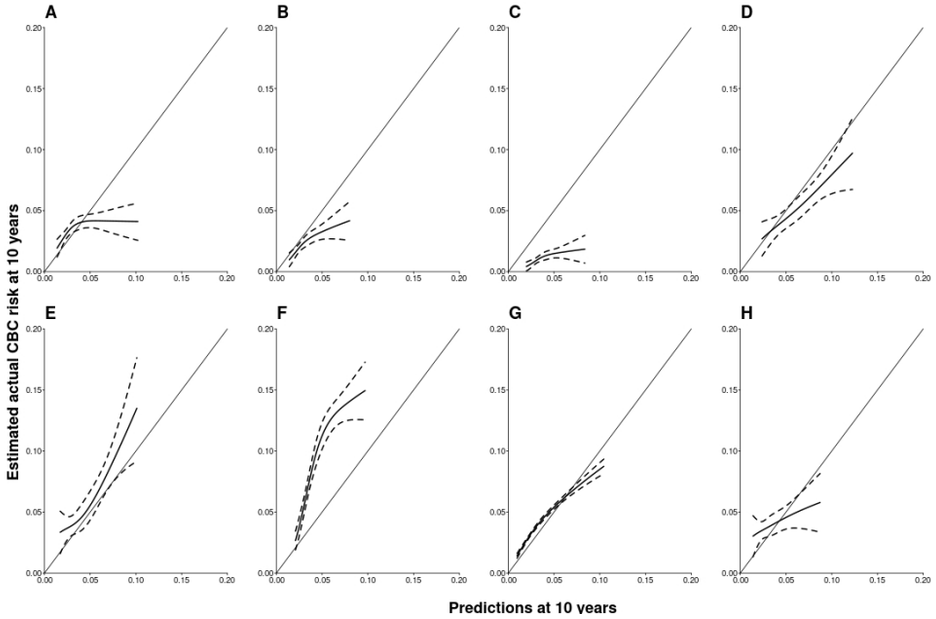
Supplementary Figure 1: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 5 years for the PredictCBC-2.0A model.
The x-axis represents the predicted cumulative incidence of contralateral breast cancer estimated by PredictCBC-2.0A model at 5 years and the y-axis the estimated actual contralateral breast cancer risk at 5 years. The black lines indicate the calibration of predicted values using an three-knot restricted cubic spline. Dashed black lines indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Europe – Other; Panel B: Europe – Scandinavia; Panel C: Europe – UK ; Panel D: Netherlands – BOSOM; Panel E: Netherlands – EMC; Panel F: Netherlands – HEBON; Panel G: Netherlands – NCR; Panel H: US and Australia.



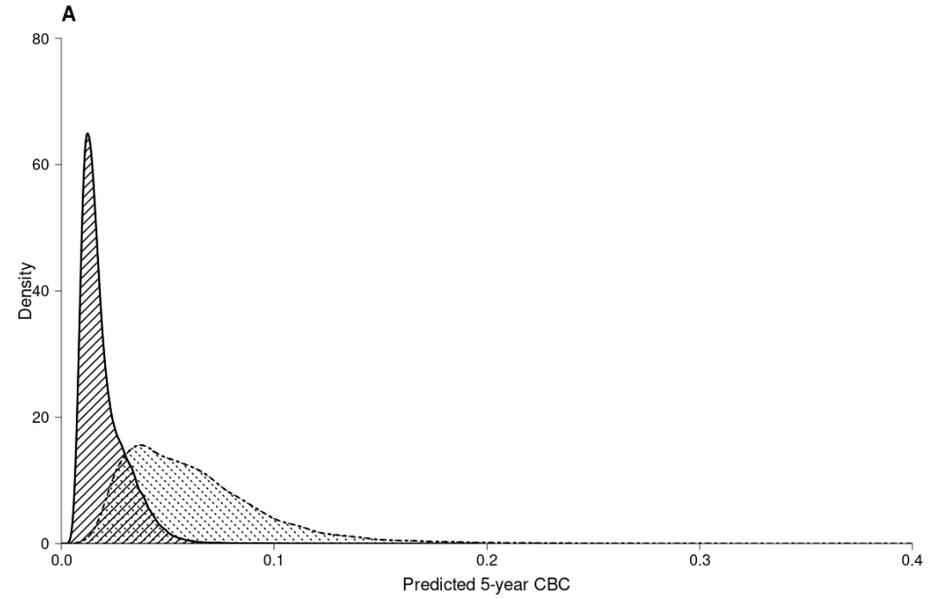
Supplementary Figure 2: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 10 years for the PredictCBC-2.0A model.
The x-axis represents the predicted cumulative incidence of contralateral breast cancer estimated by PredictCBC-2.0A model at 10 years and the y-axis the estimated actual contralateral breast cancer risk at 10 years. The black lines indicate the calibration of predicted values using a three-knot restricted cubic spline. Dashed black lines indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Europe – Other; Panel B: Europe – Scandinavia; Panel C: Europe – UK; Panel D: Netherlands – BOSOM; Panel E: Netherlands – EMC; Panel F: Netherlands – HEBON; Panel G: Netherlands – NCR; Panel H: US and Australia.



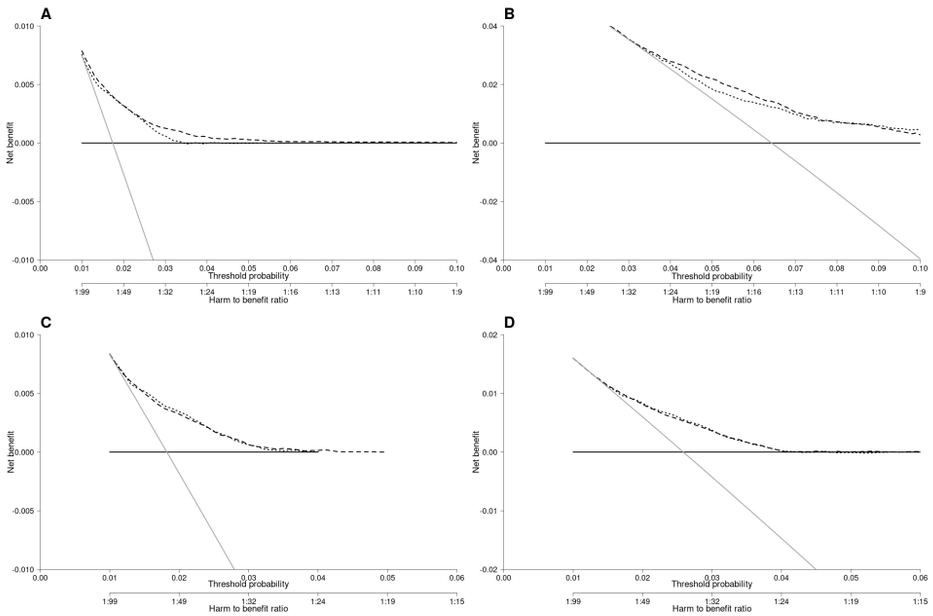
Supplementary Figure 3: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 5 years for the PredictCBC-2.0B model.
The x-axis represents the predicted cumulative incidence of contralateral breast cancer estimated by PredictCBC-2.0B model at 5 years and the y-axis the estimated actual contralateral breast cancer risk at 5 years. The black lines indicate the calibration of predicted values using a three-knot restricted cubic spline. Dashed black lines indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Europe – Other; Panel B: Europe – Scandinavia; Panel C: Europe – UK; Panel D: Netherlands – BOSOM; Panel E: Netherlands – EMC; Panel F: Netherlands – HEBON; Panel G: Netherlands – NCR; Panel H: US and Australia.



Supplementary Figure 4: Visual assessment of calibration through calibration plots in the internal-external cross-validation at 10 years for the PredictCBC-2.0B model.
The x-axis represents the predicted cumulative incidence of contralateral breast cancer estimated by PredictCBC-2.0B model at 10 years and the y-axis the estimated actual contralateral breast cancer risk at 10 years. The black lines indicate the calibration of predicted values using an three-knot restricted cubic spline. Dashed black lines indicate the 95% confidence intervals. The dashed gray line indicates perfect overall calibration. Each panel indicates a validation in one of the datasets. Panel A: Europe – Other; Panel B: Europe – Scandinavia; Panel C: Europe – UK; Panel D: Netherlands – BOSOM; Panel E: Netherlands – EMC; Panel F: Netherlands – HEBON; Panel G: Netherlands – NCR; Panel H: US and Australia.



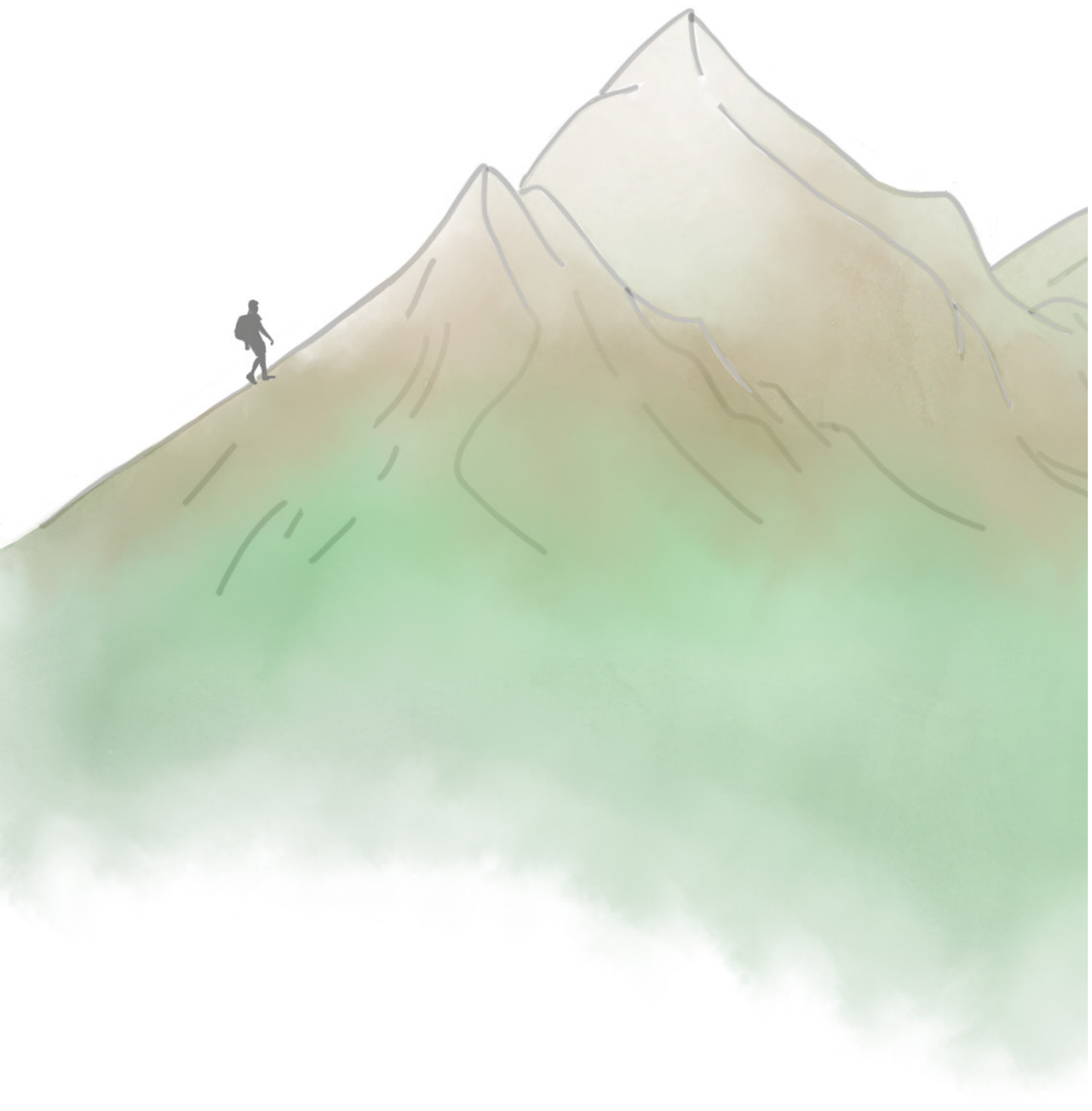
Supplementary Figure 5: Density distribution of 5-year predicted contralateral breast cancer using PredictCBC-2.0 models. **a** Density distribution of 5-year predicted contralateral breast cancer absolute risk using PredictCBC-2.0A within non-carriers (area with black solid lines) and *BRCA1/2* mutation carriers (area with black dashed lines). **b** Density distribution of 5-year predicted contralateral breast cancer absolute risk using PredictCBC-2.0B within patients without (first degree) family history (area with black solid lines) and patients with (first degree) family history (area with black dashed lines).



Supplementary Figure 6. Decision curve analysis at 5 years for the contralateral breast cancer risk models (PredictCBC and PredictCBC-2.0) including *BRCA* mutation information. **a** The decision curve to determine the net benefit of the estimated 5-year predicted contralateral breast cancer (CBC) cumulative incidence for patients without a *BRCA1/2* gene mutation using PredictCBC-1A (dotted black line) and PredictCBC-2.0A (dashed black line) compared to not treating any patients with contralateral preventive mastectomy (CPM) (black solid line). **b** The decision curve to determine the net benefit of the estimated 5-year predicted CBC cumulative incidence for *BRCA1/2* mutation carriers using PredictCBC-1A (dotted black line), PredictCBC-2.0A (dashed black line) versus treating (or at least counseling) all patients (gray solid line). **c** The decision curve to determine the net benefit of the estimated 5-year predicted CBC cumulative incidence for patients without (first-degree) family history using PredictCBC-1B (dotted black line), PredictCBC-2.0B (dashed black line) compared to not treating any patients with CPM (black solid line). **d** The decision curve to determine the net benefit of the estimated 5-year predicted CBC cumulative incidence for patients with (first-degree) family history using PredictCBC-1B (dotted black line), PredictCBC-2.0B (dashed black line) versus treating (or at least counseling) all patients (gray solid line). The y-axis measures net benefit, which is calculated by summing the benefits (true positives, i.e., patients with a CBC who needed a CPM) and subtracting the harms (false positives, i.e., patients with CPM who do not need it). The latter are weighted by a factor related to the relative harm of a non-prevented CBC versus an unnecessary CPM. The factor is derived from the threshold probability to develop a CBC at 10 years at which a patient would opt for CPM (e.g., 5%). The x-axis represents the threshold probability. Using a threshold probability of 5% implicitly means that CPM in 20 patients of whom one would develop a CBC if untreated is acceptable (19 unnecessary CPMs, harm to benefit ratio 1:19)

REFERENCES

1. Michailidou K, Lindstrom S, Dennis J, *et al.* Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551(7678):92-94.
2. Schmidt MK, Tollenaar RA, de Kemp SR, *et al.* Breast cancer survival and tumor characteristics in premenopausal women carrying the CHEK2*1100delC germline mutation. J Clin Oncol 2007;25(1):64-9.
3. Schmidt MK, van den Broek AJ, Tollenaar RA, *et al.* Breast Cancer Survival of BRCA1/BRCA2 Mutation Carriers in a Hospital-Based Cohort of Young Women. J Natl Cancer Inst 2017;109(8).
4. Font-Gonzalez A, Liu L, Voogd AC, *et al.* Inferior survival for young patients with contralateral compared to unilateral breast cancer: a nationwide population-based study in the Netherlands. Breast Cancer Res Treat 2013;139(3):811-9.
5. Kramer I, Schaapveld M, Oldenburg HSA, *et al.* The influence of adjuvant systemic regimens on contralateral breast cancer risk and receptor subtype. J Natl Cancer Inst In press.
6. Giardiello D, Steyerberg EW, Hauptmann M, *et al.* Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res 2019;21(1):144.
7. Pijpe A, Manders P, Brohet RM, *et al.* Physical activity and the risk of breast cancer in BRCA1/2 mutation carriers. Breast Cancer Res Treat 2010;120(1):235-44.
8. Mavaddat N, Michailidou K, Dennis J, *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019;104(1):21-34.
9. Kramer I, Hooning MJ, Mavaddat N, *et al.* Breast Cancer Polygenic Risk Score and Contralateral Breast Cancer Risk. Am J Hum Genet 2020;107(5):837-848.
10. Bouchardy C, Usel M, Verkooijen HM, *et al.* Changing pattern of age-specific breast cancer incidence in the Swiss canton of Geneva. Breast Cancer Res Treat 2010;120(2):519-23.
11. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.
12. Resche-Rigon M, White IR, Bartlett JW, *et al.* Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. Stat Med 2013;32(28):4890-905.
13. Jolani S, Debray TP, Koffijberg H, *et al.* Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. Stat Med 2015;34(11):1841-63.
14. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009;28(15):1982-98.



Chapter 5

Contralateral breast cancer risk in patients with ductal carcinoma in situ and invasive breast cancer

NPJ Breast Cancer. 2020 Nov 3;6(1):60[#]
https://www.nature.com/articles/s41523-020-00202-8

Daniele Giardiello*
Iris Kramer*
Maartje J. Hooning
Michael Hauptmann
Esther H. Lips
Elinor Sawyer
Alastair M. Thompson
Linda de Munck
Sabine Siesling
Jelle Wesseling
Ewout W. Steyerberg
Marjanka K. Schmidt

*authors contributed equally
[#]A full list of authors' affiliations appears on the journal's website

ABSTRACT

We aimed to assess contralateral breast cancer (CBC) risk in patients with ductal carcinoma in situ (DCIS) compared with invasive breast cancer (BC). Women diagnosed with DCIS (N=28,003) or stage I-III BC (N=275,836) between 1989-2017 were identified from the nationwide Netherlands Cancer Registry. Cumulative incidences were estimated, accounting for competing risks, and hazard ratios (HRs) for metachronous invasive CBC. To evaluate effects of adjuvant systemic therapy and screening, separate analyses were performed for stage I BC without adjuvant systemic therapy and by mode of first BC detection. Multivariable models including clinico-pathological and treatment data were created to assess CBC risk prediction performance in DCIS patients. The 10-year cumulative incidence of invasive CBC was 4.8% for DCIS patients (CBC=1,334). Invasive CBC risk was higher in DCIS patients compared with invasive BC overall (HR=1.10, 95% confidence interval (CI)=1.04-1.17), and lower compared with stage I BC without adjuvant systemic therapy (HR=0.87; 95%CI=0.82–0.92). In patients diagnosed ≥2011, the HR for invasive CBC was 1.38 (95%CI=1.35-1.68) after screen-detected DCIS compared with screen-detected invasive BC, and was 2.14 (95%CI=1.46-3.13) when not screen-detected. The C-index was 0.52 (95% CI=0.50-0.54) for invasive CBC prediction in DCIS patients. In conclusion, CBC risks are low overall. DCIS patients had a slightly higher risk of invasive CBC compared with invasive BC, likely explained by the risk-reducing effect of (neo)adjuvant systemic therapy among BC patients. For support of clinical decision making more information is needed to differentiate CBC risks among DCIS patients.

INTRODUCTION

Contralateral breast cancer (CBC) is the most frequent second cancer reported after first invasive breast cancer (BC)[1-3]. The cumulative incidence of invasive CBC for women following invasive BC is ~0.4% per year[4-6]. Several studies have shown a decrease in CBC incidence as a result of (neo)adjuvant systemic therapies[6-8].

Ductal carcinoma in situ (DCIS) is a potential precursor of invasive BC. The incidence of DCIS has increased substantially with widespread introduction of population-based mammography screening including digital mammography and represents 10-25% of all BC patients[9-11]. Since DCIS has an excellent prognosis with a disease-specific survival of more than 98% at 10 years[12-14], a large group of women is at risk of developing CBC.

The risk of invasive CBC for DCIS patients has not been widely investigated, but the annual risk is estimated between 0.4-0.6%[11,13,15,16]. Moreover, it is unclear if the risk of CBC is comparable between patients diagnosed with invasive BC and patients with DCIS. One study in the United States (US), using data from the Surveillance, Epidemiology, and End Results (SEER) database, found a similar relative CBC risk for DCIS patients compared to patients with invasive BC[17]. On the other hand, an indirect assessment between DCIS patients and invasive BC patients has been provided by a CBC risk prediction model developed and validated in the US, showing a higher relative CBC risk for DCIS compared with invasive BC (relative risk: 1.60, 95% confidence interval (CI)=1.42–1.93)[18,19]. The reason for a potential higher CBC risk for DCIS patients is still unclear, but might relate to the risk-reducing effect of adjuvant systemic therapy among invasive BC patients[6,20,21]. In general, relatively few DCIS patients receive adjuvant systemic therapy. In addition, CBC risks may also differ based on the mode of detection of the first BC. Previous research showed that screen-detected invasive breast tumours have a better BC-specific survival than non-screened tumours and hence receive less adjuvant systemic treatment[22].

The aim of this study was to assess the risk of developing invasive CBC in DCIS patients in direct comparison with patients diagnosed with invasive BC using a large population-based cohort of Dutch BC patients, taking age, mode of first BC detection, and (neo) adjuvant systemic therapy into account. In addition, we evaluated the CBC risk prediction performance in patients diagnosed with DCIS.

METHODS

Study population

We evaluated 323,285 patients diagnosed with in situ or invasive first BC in 1989-2017, who underwent surgery, from the Netherlands Cancer Registry (NCR) (Supplementary Figure 4). The NCR is an on-going nationwide population-based data registry of all newly diagnosed cancer patients in the Netherlands, with full coverage since 1989[23]. We excluded nine patients with first diagnosis without cytological or histological confirmation, 5,785 with stage IV BC or with incomplete staging information, 66 with squamous cell carcinoma, and 4,145 with in situ BC that was not pure DCIS (i.e. lobular, other subtype, or mixed with ductal). Follow-up for all patients started three months after the first diagnosis; therefore, 9,441 patients who had developed synchronous CBC (invasive or in situ), invasive ipsilateral BC, or died within three months after the first diagnosis were excluded.

Patient and tumour characteristics

Clinico-pathological data were provided by the NCR. After notification by the nationwide network and registry of histo- and cytopathology in the Netherlands (PALGA) and the national hospital discharge database, registration clerks of the NCR collect data directly from patients' records. Follow-up information on vital status and second cancers was complete up to January 31, 2018.

Staging was coded according to the TNM Classification of Malignant Tumours using the edition valid at the date of diagnosis, ranging from the 4th to the 8th edition[24]. If pathological stage was missing, clinical stage was used[25].

Receptor status was determined by immunohistochemistry (IHC), and was included in the NCR since 2005. Tumours were defined as estrogen receptor (ER) positive or progesterone receptor (PR) positive when >10% of the tumour cells stained positive (from 2011 the threshold was ≥10%). A tumour was defined human epidermal growth factor receptor 2/neu-receptor (HER2) positive if IHC was 3+ (strong and complete membranous expression in >10% of tumour cells) or if IHC score 2+ when additional confirmation within situ hybridization was available, but considered unknown if in situ hybridization confirmation was missing.

The NCR did not record information on *BRCA1* and *BRCA2* germline mutation status and family history.

From 2011, the NCR recorded the mode of first BC detection, i.e. if the DCIS or invasive BC was screen-detected or not detected by screening. We did not have detailed information available on the tumours not detected by screening, but these may include interval tumours, non-screen attendant, or screened outside the national program (e.g. due to family history). According to the Dutch guidelines, mammographic follow-up is similar for DCIS and invasive BC[26].

Data used in this study were included in the NCR under an opt-out regime according to Dutch legislation and codes of conduct[25]. The NCR Privacy Review Board approved this study under reference number K18.245. Data were handled in accordance with privacy regulations for medical research[25].

Statistical analyses

The primary outcome was the development of metachronous CBC, defined as an invasive BC in the contralateral breast diagnosed at least three months after the first BC diagnosis (DCIS or invasive BC). Follow-up started three months after the first BC diagnosis, and ended at date of in situ- or invasive CBC, invasive ipsilateral BC, or last date of follow-up (due to death, lost to follow-up, or end of study), whichever occurred first.

Cox proportional hazard models were performed to investigate the association of having DCIS compared with invasive BC as primary diagnosis with the cause-specific hazard of invasive CBC. We also performed analyses within situ CBC, invasive ipsilateral BC, and death as the outcome. According to the Dutch guideline, DCIS patients do not receive adjuvant systemic therapy. We evaluated the impact of adjuvant systemic therapy by comparing the invasive CBC risk between DCIS patients and patients diagnosed with stage I BC not receiving adjuvant systemic therapy (no chemotherapy, endocrine therapy, nor trastuzumab), i.e., a subgroup of patients that resembles as much as possible the DCIS patient group in terms of treatment conditions. Since hazard ratios (HRs) based on Cox regressions do not have a direct relationship with the cumulative incidence of the event of interest, we also performed competing risks regression to estimate the HRs for the subdistribution hazards of the Fine and Gray model[27,28]. In situ CBC, invasive ipsilateral BC, and death were considered as competing risks. We performed both univariable analyses and analyses adjusted for age- and year of first BC diagnosis. Since 1989, women in the Netherlands aged 50-70 have been invited for biannual screening by mammography, which was extended to women aged 75 since 1998. Based on this, we categorized age at first BC diagnosis into <50 years and ≥50 years. Based on the gradual implementation of the Dutch BC screening, we categorized year at first BC diagnosis into two periods: 1989–1998 (implementation phase) and 1999–2017 (full nationwide coverage; attendance rate is 78.8%[29] and detection rate of invasive BC 6.6 per 1000 in 2017[30]and for DCIS 0.94 per 1000 between 2004-2011[31]). We also performed our analyses stratified by mode of first BC detection. These analyses only included patients diagnosed during or after 2011 and aged 50-75 (eligible for screening).

Cumulative incidence curves of invasive CBC for DCIS patients, all invasive BC patients, and patients with stage I BC not receiving adjuvant systemic therapy were calculated considering in situ CBC, invasive ipsilateral BC, and death as competing risks. These curves were stratified by year of first BC diagnosis (1989-1998 and 1999-2017) and by age (<50 and ≥50 years).

We used joint Cox proportional hazard models[32] to investigate subtype-specific CBC risk (according to stage, grade, ER, PR, and HER2 status) in DCIS patients compared with patients with invasive BC and compared with patients with stage I BC who did not receive adjuvant systemic therapy. Each model included subtype-specific CBC (e.g. ER positive CBC, ER negative CBC, ER unknown CBC), in situ CBC, ipsilateral invasive BC, and death as possible outcomes. Since the NCR actively registered receptor status from 2005, these analyses only included patients diagnosed between 2005-2017.

Multivariable Cox regression was used to quantify the effect of clinico-pathological and treatment characteristics on CBC risk (all CBC and invasive CBC only) in DCIS patients. In addition, multivariable Fine and Gray regressions were performed to assess the association between every factor and the CBC cumulative incidence. Variables included in the models were age at first DCIS diagnosis, tumour grade, type of surgery (mastectomy or breast conserving surgery), and radiotherapy. The proportional hazard assumption of the models was assessed by examining the Schoenfeld residuals, and restricted cubic splines were used to verify whether linearity of age at first DCIS diagnosis would hold[33]. The discrimination ability of the models to identify patients developing CBC was calculated using the C-index[34]. Missing data were multiply imputed by chained equations (MICE) to avoid loss of information due to case-wise deletion causing bias and reduction in efficiency [35,36]. Multiple imputation accounts for missing data mechanisms assuming that the probability of missingness depends on the observed data namely missing at random (MAR). For every predictor with missing data, every imputation model selects predictors based on correlation structure underlying the data. Details about the imputation model are provided in Supplementary Methods.

Analyses were performed using STATA version 16.0, SAS (SAS Institute Inc., Cary, NC, USA) version 9.4, and R software version 3.5.3.[37].

RESULTS

Patient characteristics

The cohort comprised 28,003 DCIS patients (CBC=1,334) and 275,836 patients with invasive BC (CBC=12,821), including 86,481 patients with stage I BC not receiving adjuvant systemic therapy; i.e. no chemotherapy, endocrine therapy, nor trastuzumab (Table 1). The percentage of patients diagnosed with DCIS, of all BC patients diagnosed in the Netherlands, was 5.7% in the implementation phase of the mammography screening program (1989-1998) and 10.5% in the period of full national coverage (1999-2017). Median follow-up was 11.4 years.

Table 1. Patient-, tumour- and treatment characteristics of women diagnosed with ductal carcinoma in situ or invasive breast cancer

	DCIS		All invasive BC		Stage I BC without adjuvant systemic therapy[a]	
	N	%	N	%	N	%
Characteristics	**28,003**	**9.2**	**275,836**	**90.8**	**86,481**	**31.4**
Diagnosis, year						
median (range)	2009 (1989 - 2017)		2004 (1989 - 2017)		2004 (1989 - 2017)	
Age, years						
median (range)	59 (21 - 95)		59 (18 - 102)		61 (18 - 99)	
TNM stage						
0	28,003	100.0	-	-	-	-
I	-	-	120,952	43.8	86,481	100.0
II	-	-	124,883	45.3	-	-
III	-	-	30,001	10.9	-	-
Tumour grade						
I (well differentiated)	3,729	16.1	44,690	20.9	27,566	41.9
II (moderately differentiated)	7,864	33.8	95,251	44.6	28,159	42.8
III (poorly/undifferentiated)	11,639	50.1	73,581	34.5	10,036	15.3
missing	4,771	-	62,314	-	20,720	-
ER status						
positive	-	-	133,761	82.7	41,883	90.1
negative	-	-	28,075	17.3	4,598	9.9
missing	28,003	-	114,000	-	40,000	-
HER2 status						
positive	-	-	19,708	14.3	2,324	6.1
negative	-	-	118,409	85.7	35,616	93.9
missing	28,003	-	137,719	-	48,541	-
PR status						
positive	-	-	106,786	67.5	33,862	74.8
negative	-	-	51,437	32.5	11,404	25.2
missing	28,003	-	117,613	-	41,215	-
(Neo)adjuvant chemotherapy						
yes	17	0.1	91,844	33.3	-	-
no	27,986	99.9	183,992	66.7	86,481	100.0
(Neo)adjuvant endocrine therapy						
yes	102	0.4	119,394	43.3	-	-
no	27,901	99.6	156,442	56.7	86,481	100.0
(Neo)adjuvant trastuzumab						
yes	3	0.0	13,994	5.1	-	-
no	28,000	100.0	261,842	94.9	86,481	100.0

Table 1. Continued

	DCIS		All invasive BC		Stage I BC without adjuvant systemic therapy[a]	
	N	%	N	%	N	%
Characteristics	**28,003**	**9.2**	**275,836**	**90.8**	**86,481**	**31.4**
Surgery to the breast						
breast conserving surgery	16,396	60.8	142,495	53.4	58,727	70.1
mastectomy	10,571	39.2	124,530	46.6	25,023	29.9
missing	1,036	-	881	-	2,731	-
Radiation to the breast						
yes	13,128	46.9	182,226	66.1	59,354	70.1
no	14,875	53.1	93,610	33.9	27,127	31.4
Follow-up, *years*						
median (IQR)	8.7 (8.5 - 8.8)		11.8 (11.7 - 11.8)		13.5 (13.4 - 13.6)	
Cumulative incidence of invasive CBC, *%*						
5-year (95%CI)	2.4 (2.2 - 2.6)		2.0 (2.0 - 2.1)		2.9 (2.8 - 3.0)	
10-year (95%CI)	4.8 (4.6 - 5.2)		4.0 (4.0 - 4.1)		5.6 (5.4 - 5.8)	
number of invasive CBC	1,334		12,821		5,782	
Cumulative incidence of death, *%*						
5-year (95%CI)	3.8 (3.6 - 4.0)		15.0 (14.9 - 15.2)		7.8 (7.6 - 8.0)	
10-year (95%CI)	9.8 (9.4 - 10.2)		29.4 (29.2 - 29.6)		19.2 (18.9 - 19.5)	
number of death	3,340		91,797		23,899	
Cumulative incidence of ipsilateral invasive BC *%*						
5-year (95%CI)	1.6 (1.5 - 1.8)		0.1 (0.1 - 0.1)		0.2 (0.1 - 0.2)	
10-year (95%CI)	3.5 (3.3 - 3.8)		0.3 (0.2 - 0.3)		0.5 (0.4 - 0.6)	
number of ipsilateral invasive BC	920		1,471		897	
Cumulative incidence of in situ CBC, *%*						
5-year (95%CI)	1.0 (1.0 - 1.1)		0.4 (0.4 - 0.5)		0.6 (0.6 - 0.7)	
10-year (95%CI)	1.6 (1.5 - 1.8)		0.8 (0.7 - 0.8)		1.1 (1.0 - 1.2)	
number of in situ CBC	427		2,278		1,026	

Abbreviations: DCIS = ductal carcinoma in situ; BC = breast cancer; ER = estrogen-receptor; PR = progesterone-receptor; HER2 = human epidermal growth factor receptor 2; IQR = inter-quartile range; CBC = contralateral breast cancer; CI = confidence interval

[a] The 'stage I BC without adjuvant systemic therapy' group is a subset of the 'all invasive BC' group

CBC risk for patients diagnosed with DCIS and invasive BC

The 10-year cumulative incidence of invasive CBC was 4.8% (95%CI=4.6–5.2%) for DCIS patients, 4.0% (95%CI=4.0–4.1%) for all invasive BC patients, and 5.6% (95%CI=5.4–5.8%) for patients with stage I BC not receiving adjuvant systemic therapy (Table 1, Figure 1[38]). For comparison, the 10-year cumulative incidence of invasive CBC in patients diagnosed with stage I invasive BC treated with adjuvant systemic therapy was 2.9% (95%CI=2.5-3.3%). Being diagnosed with DCIS was associated with an increased risk of invasive CBC compared with invasive BC overall (HR=1.10, 95%CI=1.04–1.17), and with a lower risk when compared with stage I BC without adjuvant systemic therapy (HR=0.87, 95%CI=0.82–0.92, Table 2). Similar results were observed when using competing risk regression (Table 2).

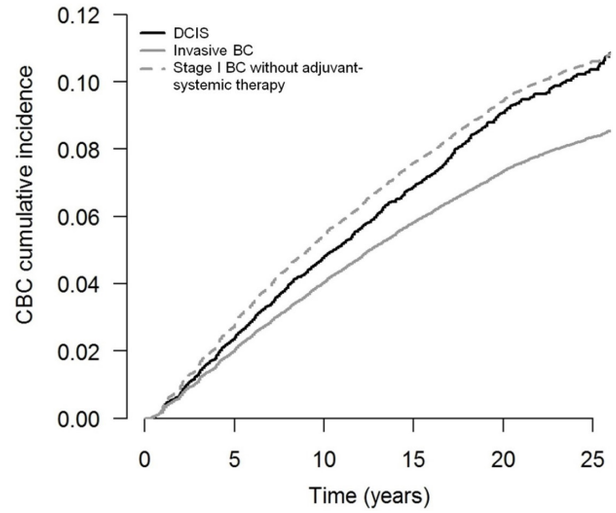


Figure 1. Cumulative incidences of invasive contralateral breast cancer (CBC) in patients diagnosed with ductal carcinoma in situ (DCIS), invasive breast cancer (BC) stage I-III, and stage I BC without (neo)adjuvant systemic therapy
The x-axis represents the time since first BC diagnosis (in years) and the y-axis the cumulative CBC incidence

Table 2. Relative subsequent contralateral breast cancer risks (invasive and in situ) after diagnosis with ductal carcinoma in situ versus invasive breast cancer using Cox and competing risk regression

Outcome(s)	Type of first BC	Cox regression		Competing risks regression	
		Unadjusted	Adjusted[a]	Unadjusted	Adjusted[a]
		HR (95% CI)	HR (95% CI)	HR[b] (95% CI)	HR[b] (95% CI)
Invasive CBC	DCIS vs invasive BC	1.08 (1.01-1.14)	1.10 (1.04-1.17)	1.22 (1.15-1.28)	1.20 (1.14-1.27)
	DCIS vs stage I BC without adjuvant systemic therapy	0.87 (0.82-0.92)	0.87 (0.82-0.92)	0.88 (0.83-0.94)	0.87 (0.82-0.93)
In situ CBC	DCIS vs invasive BC	1.92 (1.72-2.13)	1.84 (1.66-2.04)	2.12 (1.92-2.38)	1.98 (1.79-2.20)
	DCIS vs stage I BC without adjuvant systemic therapy	1.49 (1.33-1.67)	1.38 (1.22-1.55)	1.54 (1.37-1.72)	1.40 (1.25-1.58)

Abbreviations: HR = hazard ratio; CI = confidence interval; CBC = contralateral breast cancer; BC = breast cancer; DCIS = ductal carcinoma in situ

[a] Hazard ratios adjusted by age and year at first diagnosis

[b] Hazard ratios for the subdistribution hazards of the Fine and Gray model. Invasive CBC, in situ CBC, invasive ipsilateral BC, and death were taken into account as competing risks

In sensitivity analyses using different time cut-offs for metachronous CBC, results were similar. The HR for invasive CBC developed at least six months after the first BC was 1.10 (95%CI=1.04-1.17) for DCIS compared with invasive BC, and the HR was 1.09 (95%CI=1.03-1.16) using a 12-month cut-off.

The cumulative incidence of in situ CBC, death, and invasive ipsilateral BC are shown in Supplementary Figures 1-3[38]. The 10-year cumulative incidence of in situ CBC was 1.6% (95%CI=1.5–1.8%) for DCIS patients, 0.8% (95%CI=0.7–0.8%) for invasive BC patients, and 1.1% (95%CI=1.0–1.2%) for patients with stage I BC without adjuvant systemic therapy (Table 1). The risk of death was lower in DCIS patients compared to invasive BC patients (HR=0.47, 95%CI=0.45–0.49, Supplementary Table 1).

Results by age and screening (period)

Among patients who had their first BC diagnosis during the implementation phase of the national screening program (1989–1998), the risk of invasive CBC was similar in DCIS patients compared with invasive BC patients (HR=0.93, 95%CI= 0.85–1.03, Table 3, Figure 2A-C[38]). In the period of full nationwide coverage of the screening program (1999-2017), the risk of invasive CBC was higher for DCIS patients than for invasive BC patients (HR=1.19, 95%CI=1.10–1.27, Table 3, Figure 2B-D[38]). The risk of invasive CBC was lower in DCIS patients compared with patients with stage I BC not receiving adjuvant systemic therapy in both periods (1989-1998: HR=0.90; 95%CI= 0.81–1.00, and 1999-2017: HR=0.85, 95% CI: 0.79–0.91). The effects were similar stratified by age group (<50 and ≥50 years) (Table 3). The estimated 5- and 10-year cumulative incidences by age and period are shown in Supplementary Table 2.

In a subgroup of patients diagnosed during or after 2011, with information available on the mode of first BC detection, the HR of invasive CBC was 1.53 (95%CI=1.29-1.82) for DCIS patients compared with invasive BC patients, and 0.86 (95%CI=0.71-1.03) compared with patients with stage I BC without adjuvant systemic therapy (Table 4). Among all screen-detected first BCs, the HR of invasive CBC was 1.38 (95%CI=1.35-1.68) for DCIS patients compared with invasive BC patients and 0.81 (95%CI=0.66-1.00) compared with stage I BC without adjuvant systemic therapy (Table 4). When the first BC was not detected by screening, the HR of invasive CBC was 2.14 (95%CI=1.46-3.13) for DCIS patients compared to invasive BC patients and 1.04 (95%CI=0.68-1.59) compared with stage I BC without adjuvant systemic therapy (Table 4). The risk of death in patients with DCIS compared with invasive BC and stage I BC without adjuvant systemic therapy among screen-detected and not screen-detected is shown in Supplementary Table 3.

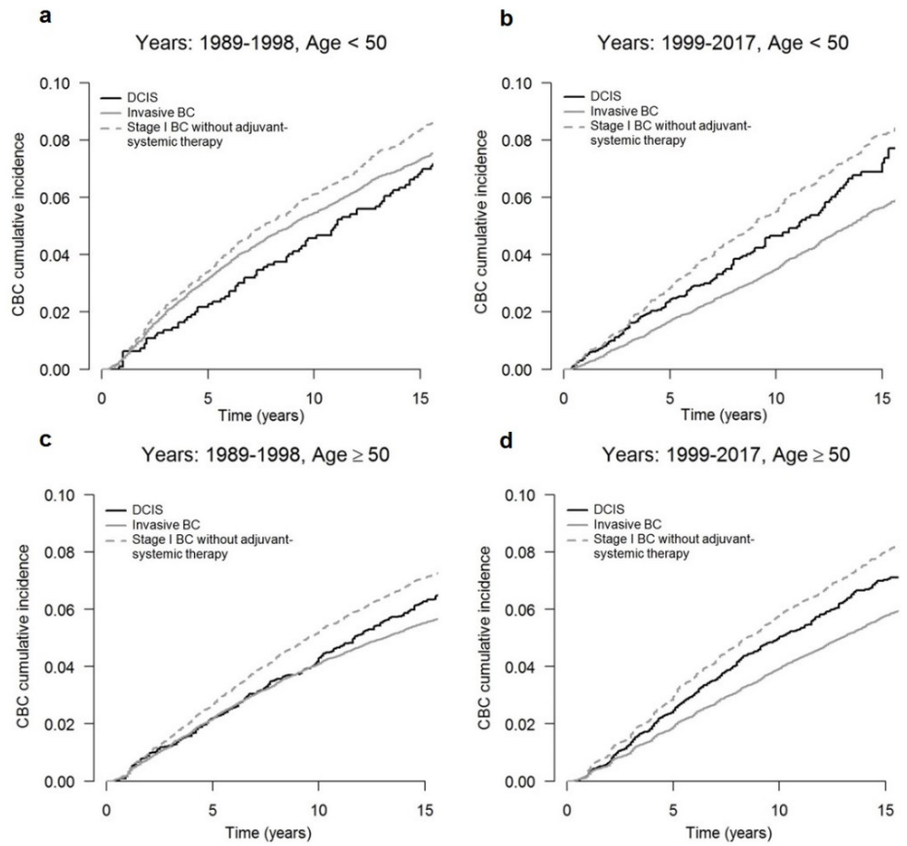


Figure 2. Cumulative incidences of invasive contralateral breast cancer (CBC) in patients diagnosed with ductal carcinoma in situ (DCIS), invasive breast cancer (BC) stage I-III, or stage I BC without (neo)adjuvant systemic therapy
Panel A) patients aged <50 years diagnosed between 1989-1998 (implementation phase Dutch mammography screening program); **Panel B)** patients aged <50 years diagnosed between 1999–2017 (full national coverage of the Dutch mammography screening program); **Panel C)** patients aged ≥50 years diagnosed between 1989–1998; **Panel D)** patients aged ≥50 years diagnosed between 1999–2017. The x-axis represents the time since first BC diagnosis (in years) and the y-axis the cumulative CBC incidence

Multivariable model

In the multivariable model, no strong predictors of CBC were identified in DCIS patients (Table 5). The C-index of the multivariable model of invasive CBC was 0.52 (standard deviation (SD=0.01) for cause-specific Cox regression; when we considered all CBC (in situ and invasive) the C-index was 0.51 (SD=0.01) (Table 5). When we performed the analyses in a subgroup of patients diagnosed during or after 2011, the C-index was 0.55 (SD=0.01) without information on the mode of first BC detection, and 0.56 (SD=0.01) with information available on the mode of first BC detection (data not shown).

Table 3. Relative risk of invasive contralateral breast cancer after ductal carcinoma in situ versus invasive breast cancer by period and age at first diagnosis using Cox and competing risks regression

					Cox regression		Competing risks regression	
	Period	Type of first BC	N	CBC events	HR	95% CI	HR[a]	95% CI
All								
	1989 - 1998	DCIS vs invasive BC	81,105	6,488	0.93	0.85 - 1.03	1.11	1.01 - 1.23
	1999 - 2017	DCIS vs invasive BC	222,734	7,667	1.19	1.10 - 1.27	1.32	1.23 - 1.41
	1989 - 1998	DCIS vs stage I BC without systemic therapy	273,383	2,696	0.90	0.81 - 1.00	0.93	0.85 - 1.04
	1999 - 2017	DCIS vs stage I BC without systemic therapy	59,098	3,086	0.85	0.79 - 0.91	0.88	0.81 - 0.94
Age < 50 years at first diagnosis[b]								
	1989 - 1998	DCIS vs invasive BC	22,084	2,292	0.94	0.83 - 1.09	1.06	0.92 - 1.22
	1999 - 2017	DCIS vs invasive BC	53,570	1,838	1.20	1.06 - 1.37	1.26	1.11 - 1.45
	1989 - 1998	DCIS vs stage I BC without systemic therapy	7,192	870	0.90	0.78 - 1.04	0.89	0.78 - 1.04
	1999 - 2017	DCIS vs stage I BC without systemic therapy	8,162	472	0.85	0.74 - 0.97	0.82	0.71 - 0.94
Age ≥ 50 years at first diagnosis[b]								
	1989 - 1998	DCIS vs invasive BC	59,021	4,196	0.92	0.83 - 1.03	1.14	1.03 - 1.26
	1999 - 2017	DCIS vs invasive BC	169,164	5,829	1.18	1.10 - 1.26	1.35	1.26 - 1.47
	1989 - 1998	DCIS vs stage I BC without systemic therapy	20,191	1,826	0.89	0.80 - 1.00	0.96	0.86 - 1.08
	1999 - 2017	DCIS vs stage I BC without systemic therapy	50,936	2,614	0.85	0.78 - 0.92	0.88	0.81 - 0.95

Abbreviations: HR = hazard ratio; CI = confidence interval; DCIS = ductal carcinoma in situ; BC = breast cancer

[a] Hazard ratios for the subdistribution hazards of the Fine and Gray model. Invasive CBC, in situ CBC, invasive ipsilateral BC, and death were taken into account as competing risks

[b] Results were based on interaction analyses including the interaction term between age, period, and type of first BC (type of first BC + age + period + age × type of first BC + period × type of first BC)

Table 4. Relative subsequent event risks after diagnosis with ductal carcinoma in situ versus invasive breast cancer by mode of first breast cancer detection for patients diagnosed between 2011-2017[a]

		Overall		By mode of first BC detection[b]		
Outcome	Type of first BC	Cox regression	Competing risks regression		Cox regression	Competing risks regression
		HR (95% CI)[c]	HR[c,d](95% CI)		HR[c] (95% CI)	HR[c,d] (95% CI)
Invasive CBC	DCIS vs invasive BC (n=62,533, events=763)	1.53 (1.29-1.82)	1.55 (1.30-1.85)	screen-detected[e]	1.38 (1.35-1.68)	1.38 (1.13-1.69)
				not screen-detected[e]	2.14 (1.46-3.13)	2.20 (1.50-3.22)
	DCIS vs stage I BC without systemic therapy (n=27,288, events=519)	0.86 (0.71-1.03)	0.86 (0.71-1.03)	screen-detected[e]	0.81 (0.66-1.00)	0.81 (0.65-1.00)
				not screen-detected[e]	1.04 (0.68-1.59)	1.05 (0.68-1.60)
In situ CBC	DCIS vs invasive BC (n=62,533, events=250)	1.99 (1.51-2.63)	2.00 (1.52-2.65)	screen-detected[e]	1.75 (1.26-2.45)	1.75 (1.26-2.45)
				not screen-detected[e]	3.41 (1.98-5.87)	3.46 (2.01-5.97)
	DCIS vs stage I BC without systemic therapy (n=27,288, events=146)	1.51 (1.08-2.10)	1.51 (1.08-2.10)	screen-detected[e]	1.40 (0.96-2.06)	1.41 (0.96-2.06)
				not screen-detected[e]	2.23 (1.14-4.39)	2.25 (1.15-4.41)

Abbreviations: BC = breast cancer; HR = hazard ratio; CI = confidence interval; CBC = contralateral breast cancer; DCIS = ductal carcinoma in situ

[a] The analyses were performed in all patients diagnosed between 2011-2017, since from 2011 we had virtually complete information on the mode of first BC detection

[b] Results were based on interaction analyses including the interaction term between mode of first BC detection and type of first BC (type of first BC + mode of first BC detection + mode of first BC detection × type of first BC)

[c] Adjusted for age at first BC diagnosis

[d] Hazard ratios for the subdistribution hazards of the Fine and Gray model. Invasive CBC, in situ CBC, invasive ipsilateral BC, and death were taken into account as competing risks

[e] Not screen-detected includes interval tumours, non-screen attendant, or screened outside the national program

Table 5. Relative risks of invasive and in situ contralateral breast cancer after diagnosis with ductal carcinoma in situ using multivariable Cox and competing risk regression models

Outcome	Invasive CBC				Invasive and in situ CBC			
	Cox regression		Competing risk regression		Cox regression		Competing risk regression	
	HR	95% CI	HR[a]	95% CI	HR	95% CI	HR[a]	95% CI
Age (years)	1.01[b]	0.93 - 1.10	0.78[c]	0.69 - 0.89	0.93[b]	0.87 - 1.00	0.71[c]	0.63 - 0.81
Tumour grade								
Moderately differentiated versus well differentiated	0.93	0.78 - 1.12	0.94	0.79 - 1.12	0.99	0.85 - 1.16	0.99	0.85 - 1.16
Poorly differentiated versus well differentiated	0.92	0.76 - 1.10	0.93	0.77 - 1.11	0.94	0.81 - 1.09	0.94	0.81 - 1.09
Surgery (Mastectomy versus BCS)	0.96	0.80 - 1.16	1.00	0.83 - 1.21	1.08	0.92 - 1.26	1.13	0.96 - 1.32
Radiotherapy to the breast (yes versus no)	1.11	0.94 - 1.32	1.12	0.94 - 1.33	1.12	0.97 - 1.30	1.14	0.98 - 1.32
Baseline failure-free probability at 10 years[d]	0.051		0.044[e]		0.068		0.057[e]	
C-index (SD)	0.520 (0.01)		0.515 (0.01)		0.513 (0.01)		0.526 (0.01)	

Abbreviations: CBC = contralateral breast cancer; HR = hazard ratio; CI = confidence interval; BCS = breast conservative surgery; SD = standard deviation
[a] Hazard ratios for the subdistribution hazards of the Fine and Gray model
[b] parameterized per decade
[c] parameterized as a restricted cubic spline with three knots
[d] The baseline failure-free probability function is calculated for baseline values of the predictors included in the multivariable models
[e] Baseline failure-free probability function for the subdistribution hazard of the Fine and Gray model

Subtype-specific CBC risk

DCIS patients had a lower risk of stage IV CBC (HR=0.45, 95%CI=0.22-0.92), and higher risks of grade I invasive CBC (HR=1.55, 95%CI=1.31-1.84) and ER-positive invasive CBC (HR=1.49, 95%CI=1.33-1.66) compared with all invasive BC patients (Supplementary Table 4). Overall, the subtype-specific CBC risk in DCIS patients was comparable to patients with stage I BC not receiving adjuvant systemic therapy (Supplementary Table 4).

DISCUSSION

In this large population-based study, the 10-year cumulative incidence of invasive CBC was 4.8% for DCIS patients. The risk of developing invasive CBC was lower for DCIS patients compared with stage I BC patients not receiving adjuvant systemic therapy (HR=0.87), but the risk was slightly higher compared with all invasive BC patients (HR=1.10). A multivariable model, based on the clinical information currently available, was unable to differentiate risks of invasive CBC among DCIS patients.

The slightly higher invasive CBC risk in DCIS patients compared with all invasive BC patients may be explained by the risk-reducing effect of adjuvant systemic therapy among invasive BC patients[6,20,21]. In our previous study using NCR data[6] we showed that adjuvant endocrine therapy, chemotherapy, and trastuzumab combined with chemotherapy were associated with overall 54%, 30%, and 43% risk reductions of CBC, respectively. In our study, a large group (57%) of patients with invasive BC received (neo) adjuvant systemic therapy. According to the Dutch guidelines, DCIS patients are not offered treatment with adjuvant systemic therapy[26]. The potential influence of adjuvant systemic therapy is supported by the CBC risk evaluation in patients diagnosed with stage I BC not receiving adjuvant systemic therapy, showing a higher CBC risk in such patients than in patients diagnosed with DCIS.

To our knowledge, only one previous study in the US investigated the risk of CBC in patients with DCIS in direct comparison with patients diagnosed with invasive BC using SEER data[17]. They found a similar CBC risk (including in situ and invasive) for invasive ductal BC in comparison with DCIS, with a relative risk of 0.98 (95%CI= 0.90–1.06). However, that analysis was based on an earlier, largely pre-screening, period (1973-1996), and lacked information on adjuvant systemic therapy use. Previous studies examining cohorts of DCIS patients have reported a subsequent annual invasive CBC risk of 0.4 to 0.6%[13,15,16], comparable to our finding.

When analyses were restricted to patients with information available on the mode of first BC detection, trends were similar overall. However, the higher CBC risk for DCIS

patients compared with invasive BC was more pronounced within the not screen-detected BC group compared with the screen-detected BC group. Tumours not detected by screening could be interval tumours or those arising in women not attending for screening. Certainly, invasive interval tumours tend to be more aggressive than screen-detected BCs and hence receive more often adjuvant systemic treatment[22].

We observed that the invasive CBCs developed within the DCIS group were less aggressive than the invasive CBCs developed after invasive first BC, i.e. more ER-positive, and lower tumour stage and grade. This may be explained by underlying etiological factors and/or be related to the use of adjuvant systemic therapy among invasive BC patients. Studies have shown that adjuvant systemic therapy influences subtype-specific CBC risk, e.g. endocrine therapy strongly reduces the risk of developing ER-positive CBC, but not ER-negative CBC[6,21]. This is supported by our subgroup analyses in patients with stage I BC not receiving adjuvant systemic therapy, who tended to develop similar CBC subtypes compared with DCIS patients.

The main strength of this study was the use of a large population-based nationwide cohort of DCIS and invasive BC patients, with complete follow-up on CBC over a long period. The NCR did not have follow-up information on distant metastases for all years included and therefore we could not take distant metastasis as a competing event into account. However, in the years where we had information on distant metastases (2003-2006), the median survival was 1.1 years and the 5-year overall survival after distant metastasis was fairly poor (6%). This indicates that death could be used as a proxy for distant metastasis. Since we had complete information on death (as a competing event), we do not expect that the lack of information on distant metastases has led to an underestimation of the CBC risk. We also did not have information available about contralateral prophylactic mastectomy (CPM), which may have resulted in an underestimation of the CBC risk and may not have had equal uptake in all groups. According to Dutch guidelines [26] only women carrying a *BRCA1 or BRCA2* germline mutation are advised to undergo a contralateral preventive mastectomy, since their CBC risk is high with an estimated 10-year risk of ~10-20%[39,40] Unfortunately, information about *BRCA1* and *BRCA2* mutation was lacking. However, we do not expect that this missing information importantly influenced the results since only 1-2% of the DCIS population[41], and 3-5% of the invasive BC population[39,42] will be *BRCA1* or *BRCA2* mutation carriers. Finally, less than 1% of the DCIS patients were not treated according to the Dutch guideline since they received adjuvant chemotherapy, endocrine therapy, and/or trastuzumab. However, since this number is low, we do not expect that this affected our results.

Despite low CBC risks, the use of CPM has increased in recent years, both in patients diagnosed with invasive BC and in patients diagnosed with DCIS, especially in the US[14,43]. Therefore, a need of individualized CBC risk prediction may be as important for patients diagnosed with DCIS as for patients with invasive BC. Currently, CBC risk prediction models have been developed and validated for patients with invasive BC, but these models may not be appropriate for DCIS patients since most of the information available for invasive BC is not routinely collected in DCIS[18,19,44,45]. In our study, we had limited information on biological characteristics of DCIS, e.g. no information on receptor subtypes, and our multivariable model was therefore unable to differentiate CBC risk among DCIS patients. So, based on the clinical information currently available, CBC risk prediction in DCIS patients is insufficiently robust to be clinically actionable. More biological knowledge is needed to improve CBC prediction in DCIS patients.

Based on the results of this study we do not suggest starting treating DCIS patients with adjuvant systemic therapy to prevent CBC since the absolute invasive CBC risk is low. To facilitate patients and physicians in decision making, a comprehensive risk prediction model specifically developed for patients with DCIS would be desirable, including information on genetic, clinical, and lifestyle factors.

Article information

Data availability statement

The datasets generated and/or analysed during the current study are not publicly available, as the study has used external data from the Netherlands Cancer Registry. The datasets will be made available from the Netherlands Cancer Registry upon reasonable request (data request study number K18.245). To apply for data access, please visit https://www.iknl.nl/en/ncr/apply-for-data. The datasets that support figures 1 and 2, and supplementary figures 1-3, are publicly available in the figshare repository, in the following data record: https://doi.org/10.6084/m9.figshare.12982424 [23].

Code availability statement

The codes developed during this study are available upon reasonable request. Analyses were performed using STATA version 16.0, SAS (SAS Institute Inc., Cary, NC, USA) version 9.4, and R software version 3.5.3.

Acknowledgements

The authors thank the registration team of the Netherlands Comprehensive Cancer Organization (IKNL) for the collection of data for the Netherlands Cancer Registry (NCR) as well as IKNL staff for scientific advice. We thank all patients whose data we used for this study and the clinicians who treated these patients. This work was supported by the Alpe d'HuZes/Dutch Cancer Society (KWF Kankerbestrijding) [grant number A6C/6253] and by Cancer Research UK/KWF Kankerbestrijding [grant numbers C38317, A24043].

The funders had no role in the design of the study, the statistical analyses, interpretation of the data, and writing of the manuscript.

Author contributions

The data used for this study were derived from by the Netherlands Cancer Registry. MKS designed the study; IK prepared and coded the data for analysis; DG performed the statistical analyses; IK, DG, MKS interpreted the results and drafted the first version of the manuscript; all other authors contributed to the interpretation of the results and revisions of the manuscript. DG and IK shared co-first authorship. All authors approved the final manuscript.

Competing interests

The authors have no conflicts of interest.

REFERENCES

1 Evans, H. S. *et al.* Incidence of multiple primary cancers in a cohort of women diagnosed with breast cancer in southeast England. *Br J Cancer* **84**, 435-440, doi:10.1054/bjoc.2000.1603 (2001).

2 Soerjomataram, I. *et al.* Primary malignancy after primary female breast cancer in the South of the Netherlands, 1972-2001. *Breast cancer research and treatment* **93**, 91-95, doi:10.1007/s10549-005-4016-2 (2005).

3 Brenner, H., Siegle, S., Stegmaier, C. & Ziegler, H. Second primary neoplasms following breast cancer in Saarland, Germany, 1968-1987. *European journal of cancer (Oxford, England : 1990)* **29a**, 1410-1414, doi:10.1016/0959-8049(93)90013-6 (1993).

4 Portschy, P. R. *et al.* Perceptions of Contralateral Breast Cancer Risk: A Prospective, Longitudinal Study. *Ann Surg Oncol* **22**, 3846-3852, doi:10.1245/s10434-015-4442-2 (2015).

5 Hartman, M. *et al.* Genetic implications of bilateral breast cancer: a population based cohort study. *Lancet Oncol* **6**, 377-382, doi:10.1016/S1470-2045(05)70174-1 (2005).

6 Kramer, I. *et al.* The Influence of Adjuvant Systemic Regimens on Contralateral Breast Cancer Risk and Receptor Subtype. *J Natl Cancer Inst* **111**, 709-718, doi:10.1093/jnci/djz010 (2019).

7 Prater, J., Valeri, F., Korol, D., Rohrmann, S. & Dehler, S. Incidence of metachronous contralateral breast cancer in the Canton of Zurich: a population-based study of the cancer registry. *J Cancer Res Clin Oncol* **142**, 365-371, doi:10.1007/s00432-015-2031-1 (2016).

8 Nichols, H. B., Berrington de Gonzalez, A., Lacey, J. V., Jr., Rosenberg, P. S. & Anderson, W. F. Declining incidence of contralateral breast cancer in the United States from 1975 to 2006. *J Clin Oncol* **29**, 1564-1569, doi:10.1200/JCO.2010.32.7395 (2011).

9 Netherlands Cancer Registry (NCR). *Survival and prevalence of cancer*, <https://www.cijfersoverkanker.nl> (2016).

10 Ernster, V. L. *et al.* Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst* **94**, 1546-1554 (2002).

11 Elshof, L. E. *et al.* Subsequent risk of ipsilateral and contralateral invasive breast cancer after treatment for ductal carcinoma in situ: incidence and the effect of radiotherapy in a population-based cohort of 10,090 women. *Breast Cancer Res Treat* **159**, 553-563, doi:10.1007/s10549-016-3973-y (2016).

12 Mariotti, C. Ductal Carcinoma in Situ of the Breast. *Springer International Publishing* (2018).

13 Miller, M. E. *et al.* Contralateral Breast Cancer Risk in Women with Ductal Carcinoma In Situ: Is it High Enough to Justify Bilateral Mastectomy? *Ann Surg Oncol* **24**, 2889-2897, doi:10.1245/s10434-017-5931-2 (2017).

14 Tuttle, T. M. *et al.* Increasing rates of contralateral prophylactic mastectomy among patients with ductal carcinoma in situ. *J Clin Oncol* **27**, 1362-1367, doi:10.1200/JCO.2008.20.1681 (2009).

15 Falk, R. S., Hofvind, S., Skaane, P. & Haldorsen, T. Second events following ductal carcinoma in situ of the breast: a register-based cohort study. *Breast Cancer Res Treat* **129**, 929-938, doi:10.1007/s10549-011-1531-1 (2011).

16 Claus, E. B., Stowe, M., Carter, D. & Holford, T. The risk of a contralateral breast cancer among women diagnosed with ductal and lobular breast carcinoma in situ: data from the Connecticut Tumor Registry. *Breast* **12**, 451-456 (2003).

17 Gao, X., Fisher, S. G. & Emami, B. Risk of second primary cancer in the contralateral breast in women treated for early-stage breast cancer: a population-based study. *International journal of radiation oncology, biology, physics* **56**, 1038-1045 (2003).

18 Chowdhury, M., Euhus, D., Onega, T., Biswas, S. & Choudhary, P. K. A model for individualized risk prediction of contralateral breast cancer. *Breast Cancer Res Treat* **161**, 153-160, doi:10.1007/s10549-016-4039-x (2017).

19 Chowdhury, M. *et al.* Validation of a personalized risk prediction model for contralateral breast cancer. *Breast Cancer Res Treat* **170**, 415-423, doi:10.1007/s10549-018-4763-5 (2018).

20 Akdeniz, D. *et al.* Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. *Breast* **44**, 1-14, doi:10.1016/j.breast.2018.11.005 (2018).

21 Langballe, R. *et al.* Systemic therapy for breast cancer and risk of subsequent contralateral breast cancer in the WECARE Study. *Breast Cancer Res* **18**, 65, doi:10.1186/s13058-016-0726-0 (2016).

22 Mook, S. *et al.* Independent prognostic value of screen detection in invasive breast cancer. *J Natl Cancer Inst* **103**, 585-597, doi:10.1093/jnci/djr043 (2011).

23 Font-Gonzalez, A. *et al.* Inferior survival for young patients with contralateral compared to unilateral breast cancer: a nationwide population-based study in the Netherlands. *Breast cancer research and treatment* **139**, 811-819, doi:10.1007/s10549-013-2588-9 (2013).

24 Brierley, J. D., Gospodarowicz, M. K. & Wittekind, C. *TNM classification of malignant tumours*. 8th Editor edn, (2017).

25 Foundation Federation of Dutch Medical Scientific Societies. Human Tissue and Medical Research: Code of Conduct for responsible use. (2011).

26 Oncoline. *Borstkanker. Landelijke richtlijn, Versie: 2.0 (August 2020 data last accessed)* < https://www.oncoline.nl/> (

27 Latouche, A., Allignol, A., Beyersmann, J., Labopin, M. & Fine, J. P. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *J Clin Epidemiol* **66**, 648-653, doi:10.1016/j.jclinepi.2012.09.017 (2013).

28 Van Der Pas, S., Nelissen, R. & Fiocco, M. Different competing risks models for different questions may give similar results in arthroplasty registers in the presence of few events. *Acta Orthop* **89**, 145-151, doi:10.1080/17453674.2018.1427314 (2018).

29 RIVM. *Breast Cancer screening program; facts and figures (May 2020, date last accessed)*, <https://www.rivm.nl/en/breast-cancer-screening-programme/background/facts-and-figures> (

30 IKNL. *National evaluation of breast cancer screening in the Netherlands 2017/2018 (August 2020, date last accessed)*, <https://www.iknl.nl/getmedia/8b019b63-0eb1-4afa-a824-31c4d10cc86e/Breast_cancer_screening_in_the_Netherlands_2017-2018_en.pdf> (

31 Sankatsing, V. D. V. *et al.* Detection and interval cancer rates during the transition from screen-film to digital mammography in population-based screening. *BMC Cancer* **18**, 256, doi:10.1186/s12885-018-4122-2 (2018).

32 Xue, X. *et al.* A comparison of the polytomous logistic regression and joint cox proportional hazards models for evaluating multiple disease subtypes in prospective cohort studies. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **22**, 275-285, doi:10.1158/1055-9965.epi-12-1050 (2013).

33 Harrell, F. E., Jr. Regression Modeling Strategies with applications to linear models, logistic and ordinal regression, and survival analysis. *Springer Series in Statistics* **2nd edition** (2015).

34 Koziol, J. A. & Jia, Z. The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data. *Biom J* **51**, 467-474, doi:10.1002/bimj.200800228 (2009).

35 Van Buuren, S. *Flexible imputation of missing data*. Second edn, (Chapman and Hall/CRC, 2018).

36 Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol* **110**, 63-73, doi:10.1016/j.jclinepi.2019.02.016 (2019).

37 R: A Language and Environment for Statistical Computing (R: Foundation for Statistical Computing, 2020).

38 Giardiello, D. *et al.* Data and metadata supporting the published article: Contralateral breast cancer risk in patients with ductal carcinoma in situ and invasive breast cancer. *figshare*, doi:https://doi.org/10.6084/m9.figshare.12982424 (2020).

39 van den Broek, A. J. *et al.* Impact of Age at Primary Breast Cancer on Contralateral Breast Cancer Risk in BRCA1/2 Mutation Carriers. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **34**, 409-418, doi:10.1200/jco.2015.62.3942 (2016).

40 Kuchenbaecker, K. B. *et al.* Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* **317**, 2402-2416, doi:10.1001/jama.2017.7112 (2017).

41 Claus, E. B., Petruzella, S., Matloff, E. & Carter, D. Prevalence of BRCA1 and BRCA2 mutations in women diagnosed with ductal carcinoma in situ. *JAMA* **293**, 964-969, doi:10.1001/jama.293.8.964 (2005).

42 Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *Journal of mammary gland biology and neoplasia* **9**, 221-236, doi:10.1023/B:JOMG.0000048770.90334.3b (2004).

43 Murphy, J. A., Milner, T. D. & O'Donoghue, J. M. Contralateral risk-reducing mastectomy in sporadic breast cancer. *Lancet Oncol* **14**, e262-269, doi:10.1016/S1470-2045(13)70047-0 (2013).

44 Basu, N. N., Ross, G. L., Evans, D. G. & Barr, L. The Manchester guidelines for contralateral risk-reducing mastectomy. *World J Surg Oncol* **13**, 237, doi:10.1186/s12957-015-0638-y (2015).

45 O'Donnell, M. Estimating Contralateral Breast Cancer Risk. *Current Breast Cancer Reports* **10**, 91-97 (2018).

SUPPLEMENTARY MATERIAL

Supplementary Methods

Multiple imputation of missing values

The predictors for contralateral breast cancer with missing values among patients diagnosed with ductal carcinoma in situ (DCIS) were type of surgery to the breast (3.7%) and tumour grade (17.0%). We used five imputed datasets based on the multiple imputation chained equations (MICE) using 50 iterations. The visit sequence of the variables was in ascending order of the number of missing values. This technique improves the accuracy and the statistical power assuming missing is at random (MAR) [1]. In the imputation procedure, we also used the year of DCIS diagnosis since this information provides a better correlation structure among covariates used as predictors in the imputation model. Continuous, binary and multiple categorical variables were imputed using predictive mean matching, binary and multinomial logistic regression, respectively. Time-to-event outcome defined as time to contralateral breast cancer, time to death, and time to ipsilateral breast cancer were included in the imputation process through the Nelson-Aalen cumulative hazard estimator[2]. For every variable with missing data, every imputation model selects predictors based on correlation structure underlying the data. We used the R package mice (version 3.6.0) to impute our data and combine the estimates using Rubin's rules.

Supplementary Table 1. Relative subsequent risks of death and invasive ipsilateral breast cancer after diagnosis with ductal carcinoma in situ versus invasive breast cancer using Cox and competing risks regression

		Cox regression		Competing risks regression	
Outcome(s)	**Type of first BC**	**Unadjusted**	**Adjusted[a]**	**Unadjusted**	**Adjusted[a]**
		HR (95% CI)	**HR (95% CI)**	**HR[b] (95% CI)**	**HR[b] (95% CI)**
Death	DCIS vs invasive BC	0.37 (0.36-0.38)	0.47 (0.45-0.49)	0.36 (0.35-0.37)	0.45 (0.44-0.47)
	DCIS vs stage I BC without adjuvant systemic therapy	0.56 (0.54-0.58)	0.71 (0.69-0.74)	0.53 (0.51-0.55)	0.68 (0.66-0.71)
	DCIS vs invasive BC	6.67 (6.25-7.14)	6.68 (6.15-7.26)	7.69 (7.14-9.09)	7.79 (7.17-8.47)
Invasive IBC	DCIS vs stage I BC without adjuvant systemic therapy	4.17 (3.85-4.54)	4.05 (3.68-4.45)	4.35 (4.00-4.76)	4.28 (3.90-4.71)

Abbreviations: HR = hazard ratio; CI = confidence interval; DCIS = ductal carcinoma in situ; BC = breast cancer; IBC = ipsilateral breast cancer

[a] Hazard ratios adjusted by age and year at first breast cancer diagnosis

[b] Hazard ratios for the subdistribution hazards of the Fine and Gray model. Invasive contralateral breast cancer, in situ contralateral breast cancer, invasive ipsilateral BC, and death were taken into account as competing risks

Supplementary Table 2. Cumulative incidence of invasive contralateral breast cancer at five and ten years in patients with ductal carcinoma in situ or invasive breast cancer by period and age at first diagnosis

	Period[a]	**Type of first BC**	**Five-year cumulative incidence (%) (95% CI)**	**Ten-year cumulative incidence (%) (95% CI)**
All	1989 - 1998	DCIS	2.2 (1.8 - 2.7)	4.4 (3.8 - 5.0)
		Invasive BC	2.5 (2.4 - 2.6)	4.5 (4.3 - 4.6)
		Stage I BC without adjuvant systemic therapy	2.9 (2.7 - 3.1)	5.5 (5.2 - 5.7)
	1999 - 2017	DCIS	2.5 (2.2 - 2.7)	5.0 (4.6 - 5.4)
		Invasive BC	1.9 (1.8 - 1.9)	3.8 (3.7 - 3.9)
		Stage I BC without adjuvant systemic therapy	3.0 (2.8 - 3.1)	5.8 (5.5 - 6.0)
Age < 50 years at first diagnosis	1989 - 1998	DCIS	2.3 (1.5 - 3.3)	4.6 (3.4 - 5.9)
		Invasive BC	3.2 (3.0 - 3.4)	5.5 (5.2 - 5.8)
		Stage I BC without adjuvant systemic therapy	3.4 (3.0 - 3.9)	6.1 (5.6 - 6.7)
	1999 - 2017	DCIS	2.4 (2.0 - 3.0)	4.7 (3.9 - 5.5)
		Invasive BC	1.7 (1.6 - 1.8)	3.5 (3.3 - 3.7)
		Stage I BC without adjuvant systemic therapy	2.9 (2.5 - 3.3)	5.5 (5.0 - 6.0)
Age ≥ 50 years at first diagnosis	1989 - 1998	DCIS	2.2 (1.8 - 2.7)	4.3 (3.7 - 5.0)
		Invasive BC	2.2 (2.1 - 2.3)	4.1 (4.0 - 4.3)
		Stage I BC without adjuvant systemic therapy	2.7 (2.5 - 2.9)	5.2 (4.9 - 5.5)
	1999 - 2017	DCIS	2.5 (2.2 - 2.7)	5.1 (4.7 - 5.4)
		Invasive BC	1.9 (1.8 - 2.0)	3.9 (3.8 - 4.0)
		Stage I BC without adjuvant systemic therapy	3.0 (2.8 - 3.1)	5.7 (5.6 - 6.0)

Abbreviations: CI = confidence interval; DCIS = ductal carcinoma in situ; BC = breast cancer

[a] The two periods were defined according to the gradual implementation of the screening program in the Netherlands: the implementation phase was between 1989 and 1998 and the full screening coverage was reached since 1999

Supplementary Table 3. Relative subsequent event risks after diagnosis with ductal carcinoma in situ versus invasive breast cancer by mode of first BC detection for patients diagnosed between 2011-2017[a]

		Overall		By mode of first BC detection[b]		
Outcome	Type of first BC	Cox regression	Competing risks regression		Cox regression	Competing risks regression
		HR (95% CI)[c]	HR[c,d] (95% CI)		HR[c] (95% CI)	HR[c,d] (95% CI)
Death	DCIS vs invasive BC (n=62,533, events=2,763)	0.48 (0.42-0.56)	0.48 (0.42-0.55)	screen-detected[e]	0.71 (0.60-0.83)	0.70 (0.60-0.83)
				not screen-detected[e]	0.33 (0.24-0.47)	0.33 (0.23-0.46)
	DCIS vs stage I BC without systemic therapy (n=27,288, events=701)	0.93 (0.79-1.09)	0.93 (0.79-1.09)	screen-detected[e]	1.04 (0.87-1.26)	1.05 (0.87-1.26)
				not screen-detected[e]	0.67 (0.46-0.98)	0.66 (0.45-0.97)
Invasive IBC	DCIS vs invasive BC (n=62,533, events=101)	5.12 (3.46-7.57)	5.17 (3.50-7.65)	screen-detected[e]	3.88 (2.46-6.14)	3.88 (2.46-6.14)
				not screen-detected[e]	10.19 (4.52-22.94)	10.42 (4.63-23.45)
	DCIS vs stage I BC without systemic therapy (n=27,288, events=83)	2.51 (1.62-3.91)	2.52 (1.62-.3.92)	screen-detected[e]	2.34 (1.41-3.88)	2.34 (1.41-3.88)
				not screen-detected[e]	3.46 (1.32-9.10)	3.48 (1.32-9.15)

Abbreviations: BC = breast cancer; HR = hazard ratio; CI = confidence interval; DCIS = ductal carcinoma in situ; IBC = ipsilateral breast cancer

[a] The analyses were performed in all patients diagnosed between 2011-2017, since from 2011 we had virtually complete information on the mode of first BC detection

[b] Results were based on interaction analyses including the interaction term between mode of first BC detection and type of first BC (type of first BC + mode of first BC detection + mode of first BC detection × type of first BC)

[c] Adjusted for age at first BC diagnosis

[d] Hazard ratios for the subdistribution hazards of the Fine and Gray model. Invasive CBC, in situ CBC, invasive ipsilateral BC, and death were taken into account as competing risks

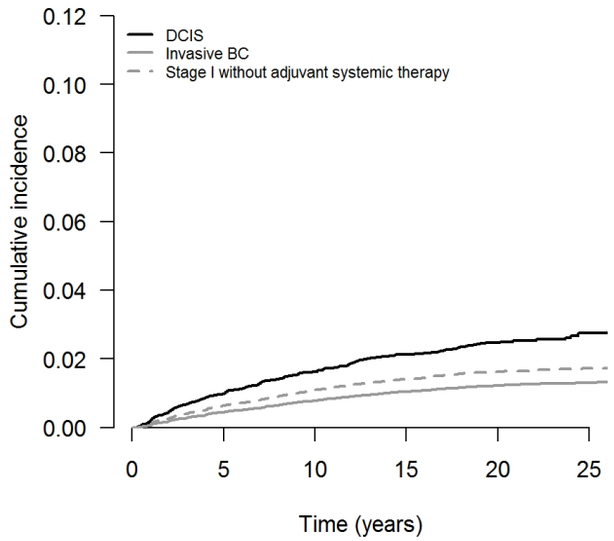
[e] Not screen-detected includes interval tumours, non-screen attendant, or screened outside the national program

Supplementary Table 4. Joint Cox regression analyses assessing subtype-specific invasive contralateral breast cancer risk for patients with ductal carcinoma in situ compared to patients with invasive breast cancer[a]

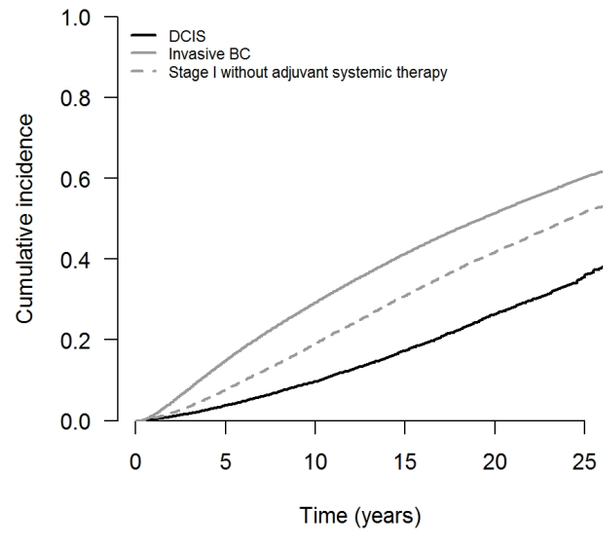
	DCIS	All invasive BC	Stage I BC without adjuvant systemic therapy	DCIS vs Invasive BC	DCIS vs Stage I BC without adjuvant systemic therapy
CBC subtypes	N	N	N	HR (95%CI)	HR (95%CI)
TNM stage					
I	330	1,957	1,084	1.35 (1.20 - 1.52)	0.74 (0.65 - 0.83)
II	146	782	342	1.50 (1.26 - 1.79)	1.04 (0.86 - 1.26)
III	40	220	78	1.46 (1.04 - 2.05)	1.26 (0.86 - 1.86)
IV	8	143	29	0.45 (0.22 - 0.92)	0.72 (0.33 - 1.58)
Tumor grade					
I (well differentiated)	154	797	518	1.55 (1.31 - 1.84)	0.72 (0.60 - 0.86)
II (moderately differentiated)	245	1,253	652	1.57 (1.37 - 1.80)	0.91 (0.79 - 1.06)
III (poorly/undifferentiated)	95	675	251	1.13 (0.91 - 1.40)	0.93 (0.73 - 1.18)
ER status					
positive	386	2,081	1,151	1.49 (1.33 - 1.66)	0.81 (0.72 - 0.91)
negative	53	471	114	0.90 (0.69 - 1.19)	1.12 (0.81 - 1.56)
PR status					
positive	314	1,560	943	1.61 (1.43 - 1.82)	0.80 (0.71 - 0.91)
negative	119	971	311	0.98 (0.81 - 1.18)	0.93 (0.75 - 1.15)
HER2 status					
positive	51	250	91	1.63 (1.21 - 2.20)	1.35 (0.96 - 1.91)
negative	375	2,200	1,133	1.36 (1.22 - 1.52)	0.80 (0.71 - 0.90)

Abbreviations: CBC = contralateral breast cancer; DCIS = ductal carcinoma in situ; BC = breast cancer; HR = hazard ratio; CI = confidence interval; ER = estrogen receptor; PR = progesterone receptor; HER2 = human epidermal growth factor receptor 2

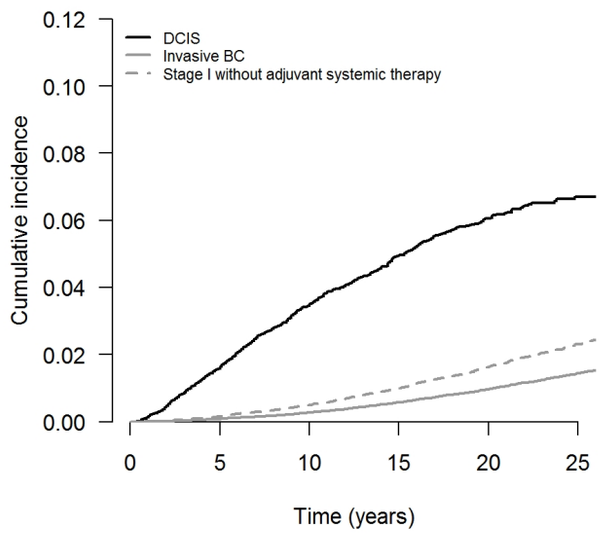
[a] The analyses were performed only in patients diagnosed between 2005-2017, since from 2005 the Netherlands Cancer Registry actively registered receptor status



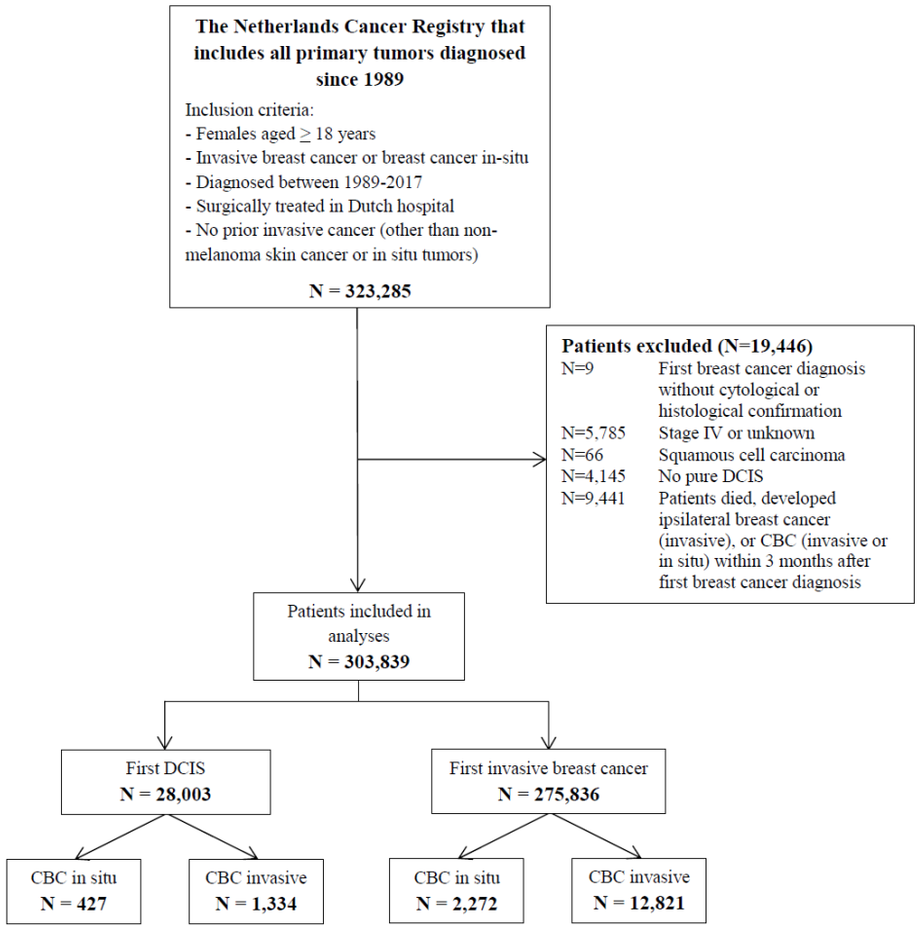
Supplementary Figure 1. Cumulative incidence of in situ contralateral breast cancer in patients diagnosed with ductal carcinoma in situ, invasive breast cancer stage I-III, and stage I breast cancer without (neo)adjuvant systemic therapy
The x-axis represents the time since the first breast cancer diagnosis (in years). The y-axis represents the cumulative incidence of in situ contralateral breast cancer. Abbreviations: DCIS = ductal carcinoma in situ; BC = breast cancer



Supplementary Figure 2. Cumulative incidence of death in patients diagnosed with ductal carcinoma in situ, invasive breast cancer stage I-III, and stage I breast cancer without (neo)adjuvant systemic therapy
The x-axis represents the time since the first breast cancer diagnosis (in years). The y-axis represents the cumulative incidence of death. Abbreviations: DCIS = ductal carcinoma in situ; BC = breast cancer



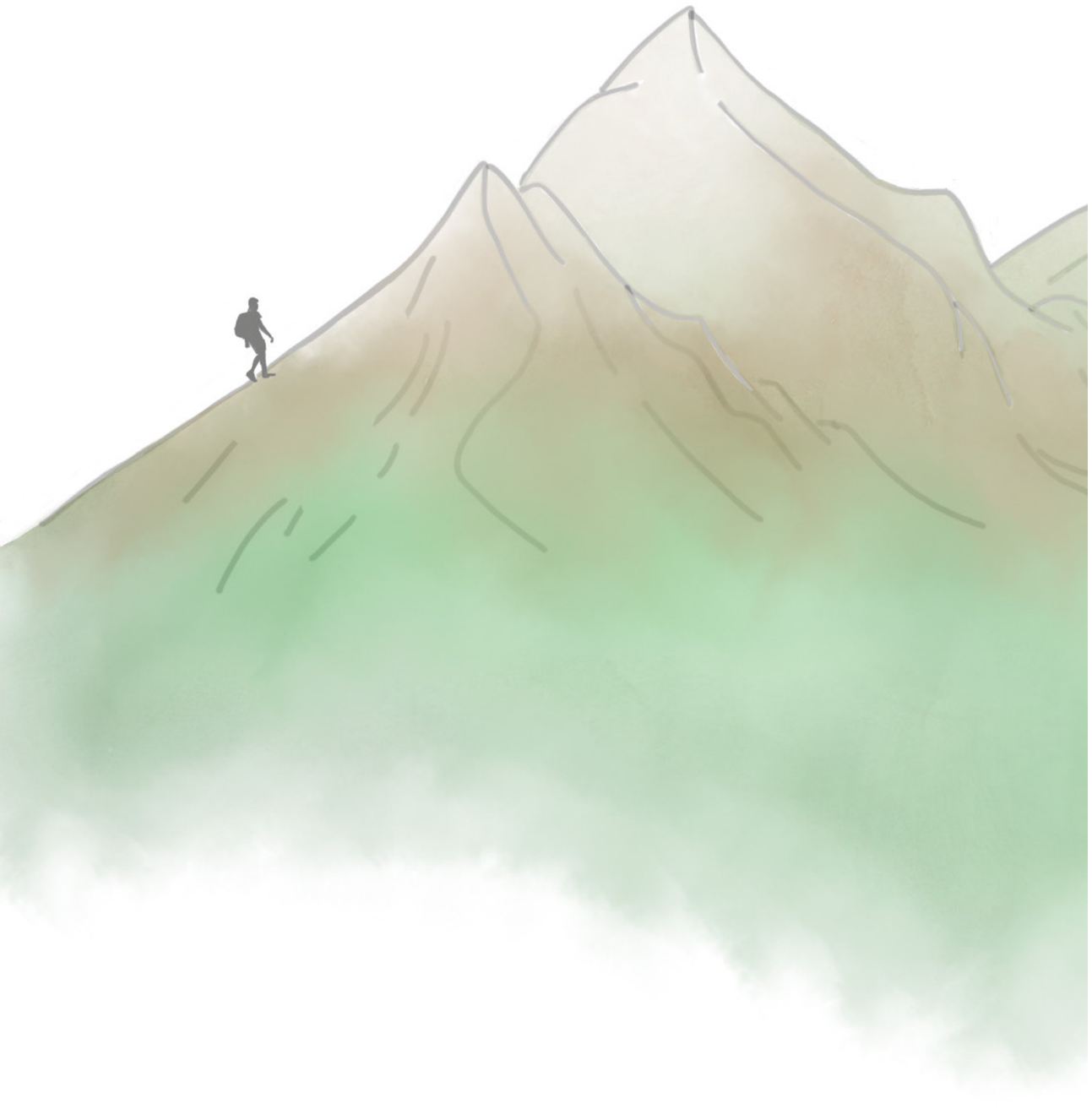
Supplementary Figure 3. Cumulative incidence of invasive ipsilateral breast cancer in patients diagnosed with ductal carcinoma in situ, invasive breast cancer stage I-III, and stage I breast cancer without (neo)adjuvant systemic therapy
The x-axis represents the time since the first breast cancer diagnosis (in years). The y-axis represents the cumulative incidence of invasive ipsilateral breast cancer. Abbreviations: DCIS = ductal carcinoma in situ; BC = breast cancer



Supplementary Figure 4. Study flowchart
Abbreviations: DCIS = ductal carcinoma in situ; CBC = contralateral breast cancer

SUPPLEMENTARY REFERENCES

1. Van Buuren, S., *Flexible imputation of missing data*. Second ed. 2018: Chapman and Hall/CRC.
2. White, I.R. and P. Royston, *Imputing missing covariate values for the Cox model.* Stat Med, 2009. **28**(15): p. 1982-98.



Chapter 6

Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for Cox proportional hazards models

Submitted for publication

David J McLernon
Daniele Giardiello
Ben van Calster
Laure Wynants
Nan van Geloven
Maarten van Smeden
Terry Therneau
Ewout W Steyerberg
on behalf of topic groups 6 and 8 of the STRATOS Initiative

ABSTRACT

Risk prediction models need thorough validation to assess their performance. Validation of models for survival outcomes poses challenges due to the censoring of observations and the varying time horizon at which predictions can be made. We aim to give a description of measures to evaluate predictions and the potential improvement in decision making from survival models based on Cox proportional hazards regression. As a motivating case study, we consider the prediction of the composite outcome of recurrence and death (the 'event') in breast cancer patients following surgery. We develop a Cox regression model with three predictors as in the Nottingham Prognostic Index in 2982 women (1275 events within 5 years of follow-up) and externally validate this model in 686 women (285 events within 5 years). The improvement in performance was assessed following the addition of circulating progesterone as a prognostic biomarker.

The model predictions can be evaluated across the full range of observed follow up times or for the event occurring by a fixed time horizon of interest. We first discuss recommended statistical measures that evaluate model performance in terms of discrimination, calibration, or overall performance. Further, we evaluate the potential clinical utility of the model to support clinical decision making. SAS and R code is provided to illustrate apparent, internal, and external validation, both for the three-predictor model and when adding progesterone.

We recommend the proposed set of performance measures for transparent reporting of the validity of predictions from survival models.

Key words: Cox regression model; survival analysis; validation; discrimination; calibration; decision analysis; STRATOS Initiative

INTRODUCTION

Prediction models for survival outcomes are important for clinicians who wish to estimate a patient's risk (i.e. probability) of experiencing a future outcome. The term 'survival' outcome is used to indicate any prognostic or time-to-event outcome, such as death, progression, or recurrence of disease. Such risk estimates for future events can support shared decision making for interventions in high-risk patients, help manage the expectations of patients, or stratify patients by disease severity for inclusion in trials.[1] For example, a prediction model for persistent pain after breast cancer surgery might be used to identify high risk patients for intervention studies.[2]

Once a prediction model has been developed it is common to first assess its performance for the underlying population. This is referred to as internal validation, which can be performed using the dataset on which the model was developed, for example by cross-validation or bootstrapping techniques.[3] External validation refers to performance in a plausibly related population, which requires an independent dataset which may differ in setting, time or place.[4, 5]

Ample guidance exists for assessing the performance of prediction models for binary outcomes, where the logistic regression model is most commonly used for model development.[6–8] Validation of a survival model poses more of a challenge due to the censoring of observation times when a patient's outcome is undetermined during the study period. If assessing 5-year survival, for instance, some subjects may have less than 5 years of follow-up without experiencing the event of interest.[3] Moreover, predictions can be evaluated over the entire range of observed follow up times or for the event occurring by a fixed time horizon of interest. The international STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (http://stratos-initiative.org) began in 2013 and aims to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies.[9]

This STRATOS article aims to provide guidance for assessing discrimination, calibration, and clinical usefulness for survival models, building on the methodological literature for survival model evaluation.[10–12] For illustration, we consider the performance of a Cox model to predict recurrence free survival (i.e. being alive and without breast cancer recurrence) at 5 years in breast cancer patients. We also describe how to assess the improvement in predictive ability and decision-making when adding a prognostic biomarker.

METHODS AND RESULTS

In the following, we discuss three key issues for the evaluation of predictions from survival prediction models. We then describe our breast cancer case study, present how we can predict survival outcomes with the Cox proportional hazards model, perform validation of predictions, and assess the potential clinical usefulness of a prediction model.

Key issues when validating a survival model

Three major issues differentiate the validation of survival models from models for binary outcomes. First, we need to decide on a time point or time range over which to assess the validation. This choice needs to be grounded in both the available data and the intended practical use of the model predictions. Altman considers a case where a model will be used for individual risk stratification in advanced pancreatic cancer patients.[13] In such a case a quite short time horizon is indicated of e.g. 18 months. Other situations with longer follow-up might use 3, 5, 10, or even 20 years.

A second issue is whether to consider prediction only up to *a fixed time point* or over an entire range of follow-up. In our case study we focus on 5 years from enrollment as the upper limit. For a cutoff of 5 years, we need to decide if only the binary outcome of whether the event occurred before or after 5 years is of interest, or also the ability to distinguish between survival of 1 and 4 years, for instance. We will give measures of performance for both settings.

A last technical issue is that estimation of the baseline survival $S_0(t)$ from the Cox model is necessary for full validation of a prediction model. However, many published reports do not provide this function (see Box 1 for further details).[10]

Description of the case study

We analysed data from patients who had primary surgery for breast cancer between 1978 and 1993 in Rotterdam.[14, 15] Patients were followed until 2007. After exclusions, 2982 patients were included in the model development cohort (Table 1). The outcome was recurrence-free survival, defined as time from primary surgery to recurrence or death. Over the maximum follow-up time of 19.3 years, 1,713 events occurred, and the estimated median potential follow-up time, calculated using the reverse Kaplan-Meier method, was 9.3 years.[16] Out of 2,982 patients, 1,275 suffered a recurrence or death within the follow-up time of interest, which was 5 years, and 126 were censored before 5 years. An external validation cohort consisted of 686 patients with primary node positive breast cancer from the German Breast Cancer Study Group,[17] where 285 suffered a recurrence or died within 5 years of follow-up, and 280 were censored before 5 years. Five-year predictions were chosen as that was the lowest median survival from the two cohorts (Rotterdam cohort, 6.7 years; German cohort, 4.9 years).

Prediction of survival outcomes

The Cox proportional hazards model is a standard for analysing survival data in biomedical settings[18] A Cox model estimates log hazard ratios, but for prediction, estimation of the baseline survival is also required. Both are needed for a full assessment of performance of a survival model in new patients (external validation, Box 1).

Box 1. The Cox proportional hazards model to make predictions for new patients

Hazard ratios express how baseline patient characteristics (or predictors) are associated with the hazard rate, that is the instantaneous rate of the event occurring at time *t*, having survived until time *t*. Mathematically, the Cox model for the hazard rate, h(t), is

$$h(t) = h_0(t)\exp(\beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \ldots + \beta_P\chi_P) = h_0(t)\exp(\mathrm{PI}),$$

where the β's are regression coefficients for the p predictors x_1 to x_p (e.g., the patient's age, disease stage, comorbidity). These regression coefficients are the log of the hazard ratios. The prognostic index, PI, represents the sum of the regression coefficients multiplied by the value of their respective predictors. The Cox model assumes that hazards for different values of a predictor are proportional during follow-up. For example, if the hazard of the event for patient A is half that of patient B at time t, the hazard ratio of 0.5 holds for these two patients at any other time point.

The baseline hazard function $h_0(t)$ is the same for all patients analogous to the intercept in linear or logistic regression models. If the primary focus of an analysis is relative risk estimation, the Cox model can be used to obtain hazard ratios without worrying about baseline hazard estimation. For estimating the risk that a patient experiences the event, i.e. absolute risk estimation, we require the baseline survival function $S_0(t)$ which is the predicted risk of survival for the patient whose predictor values are the reference categories (for categorical predictors) or zero/the mean (for continuous predictors). By integrating the hazard function from time 0 to t we obtain the cumulative hazard function, , where $h_0(t)$ is the baseline cumulative hazard function. $H_0(t)$ is then used to estimate the probability of survival up to time t , i.e. not experiencing the event up to time t:

$$S(t) = S_0(t)^{\exp\,(\beta_1\chi_1+\beta_2\chi_2+\beta_3\chi_3+\ldots+\beta_p\chi_p)} = S_0(t)^{\exp(\mathrm{PI})}$$

where $H(t) = H_0(t)\exp(PI)$, the baseline survival at time t (e.g., t = 5 years after surgery). The absolute risk of an event within t years is calculated as 1 – $S(t)$. The baseline hazard of a Cox model is often estimated non-parametrically in contrast to parametric survival models such as the accelerated failure time model.

Estimates of absolute risk are necessary for many of the performance measures discussed below. A model development study hence needs to have reported the baseline hazard function or baseline survival function, or at least survival at the time point of interest, and a specification of calculation of the PI. This is analogous to a logistic regression model to predict a binary outcome, which additionally needs reporting of a model intercept rather than only odds ratios.

Table 1: Characteristics of the breast cancer cohorts used for model development and external validation[14, 17]

Characteristic		Development cohort (n=2982, 1275 events <5 years)	Validation cohort (n=686, 285 events <5 years)
Size (mm)	≤20	1387 (46.5)	180 (26.2)
	21-50	1291 (43.3)	453 (66.0)
	>50	304 (10.2)	53 (7.7)
Number of Nodes	0	1436 (48.2)	0 (0.0)
	1 to 3	764 (25.6)	376 (54.8)
	>3	782 (26.2)	310 (45.2)
Grade of Tumour	1 or 2	794 (26.6)	525 (76.5)
	3	2188 (73.4)	161 (23.5)
Age (years: median (IQR))		54 (45 to 65)	53 (46 to 61)
Circulating progesterone (PGR, ng/mL: median (IQR))		41 (4 to 198)	33 (7 to 132)

Numbers (%) unless otherwise stated

Model development in the case study

A Cox regression model was fit to estimate recurrence free survival using three predictors: number of lymph nodes (0, 1-3, >3), tumour size (≤20mm, 21-50mm, >50mm) and pathological grade (1, 2, 3, see Table 2). Although we emphasize that it is generally poor practice to categorise continuous variables, tumour size was not available in continuous form in the dataset, and number of lymph nodes was categorised to match its form in the well-known Nottingham Prognostic Index.[1920] Since we were interested in predictons up to 5 years, we applied administrative censoring at 5 years. The Cox model assumes that hazards for different values of a predictor are proportional during follow-up. While found some evidence of non-proportional hazards (p<0.001, Grambsch and Therneau global test), we chose to ignore this violation here since it was relatively minor at graphical inspection. Furthermore, predictions made at the time of administrative censoring (5 years here) have been shown to be robust regardless of such violations.[21] The formula for the prognostic index was estimated as follows:

$$PI = 0.383 \times 1(if\ size\ is\ 21-50mm) + 0.664 \times 1(if\ size\ is > 50) + 0.360 \times 1(if\ 1\ to\ 3\ nodes) + 1.063 \times 1(if\ nodes > 3) + 0.375 \times 1(if\ grade = 3)$$

The probability of experiencing the event within 5 years can be calculated as:

$$1\text{-}S(5) = 1 - S_0(5)^{\exp(PI)} = 1 - 0.802^{\exp(PI)}$$

The baseline survival at 5 years (0.802) applies to the reference categories for the three predictors in the model (see R and SAS code in https://github.com/danielegiardiello/Prediction_performance_survival). So, a woman with a tumor size <=20mm, no nodes, and grade<3, has an estimated risk of $1 - 0.802^1 = 19.8\%$ of recurrence or breast cancer mortality within 5 years.

Table 2: Cox regression models predicting event free survival in Rotterdam breast cancer development dataset (n=2982), without and with PGR

	Without PGR	With PGR
	Hazard ratio (95% CI)	Hazard ratio (95% CI)
Size (mm)		
≤20	1	1
21-50	1.47 (1.29 to 1.67)	1.44 (1.26 to 1.63)
>50	1.94 (1.62 to 2.32)	1.90 (1.59 to 2.27)
Number of nodes		
0	1	1
1 to 3	1.43 (1.24 to 1.66)	1.46 (1.26 to 1.70)
>3	2.89 (2.52 to 3.32)	2.88 (2.51 to 3.31)
Tumour grade		
1 or 2	1	1
3	1.46 (1.27 to 1.67)	1.37 (1.19 to 1.58)
PGR (ng/ml)		
PGR1[§]		1.46[§] (1.27 to 1.68)

$$PI = 0.383 \times 1(if\ size\ is\ 21-50mm) + 0.664 \times 1(if\ size\ is > 50) + 0.360 \times 1(if\ 1\ to\ 3\ nodes) + 1.063 \times 1(if\ nodes > 3) + 0.375 \times 1(if\ grade = 3)$$

The survival at 5 years can be calculated as:

$$S(5) = 0.802^{\exp(PI)}$$

For model with PGR:

$$PI = 0.362 \times 1(if\ size\ is\ 21-50mm) + 0.641 \times 1(if\ size\ is > 50) + 0.381 \times 1(if\ 1\ to\ 3\ nodes) + 1.059 \times 1(if\ nodes > 3) + 0.317 \times 1(if\ grade = 3) - 0.003 \times PGR + 0.013 \times PGR1$$

$$where\ \ PGR1 = max\left(\frac{PGR}{61.81}, 0\right)^3 + \frac{\left(41 \times max\left(\frac{(PGR-486)}{61.81}, 0\right)^3 - 486 \times max\left(\frac{(PGR-41)}{61.81}, 0\right)^3\right)}{445}$$

The survival at 5 years can be calculated as:

$$S(5) = 0.759^{\exp(PI)}$$

[§]Since PGR was fitted as a restricted cubic spline function, it is presented as an interquartile HR to aid interpretation i.e. the hazard of mortality for the 25th percentile value (i.e. PGR=4 ng/ml) versus the hazard of mortality for the 75th percentile value (198 ng/ml).

Measures of performance

Model performance was assessed in the development dataset (apparent validation) and in the German dataset (external validation). Internal validation was assessed using the bootstrap resampling approach which provides stable estimates of performance for the population where the sample originated from. The difference between the apparent performance and the internal performance represents the "optimism" in performance of the original model (see Appendix 1 for further details).

Discrimination

A first question is how well the model predictions separate high from low risk patients:

discriminative ability. Patients with an earlier event time should exhibit a higher risk and those with later event time a lower risk.

Fixed time point discrimination

Measures that assess the prediction by a fixed time point are the similar to those for binomial outcomes. A primary issue that arises, however, is censoring in the validation data set. If we choose an evaluation time of 5 years, for instance, how are subjects who are censored before 5 years in the validation set to be assessed? For these we have a predicted risk at 5 years from the model, but do not have an observed value of the outcome at 5 years. One approach is to use inverse probability of censoring weights (IPCW), to reassign the case weights of those censored to other observations with longer follow up (see Table S1).

Uno applies such inverse weights, and this is our recommended method for assessing discrimination at a fixed time point, though many others exist.[22, 23] It assesses all pairs of patients where one experiences the event before the chosen time point and the other remains event free up to that time and calculates the proportion of those pairs for which the first mentioned patient has highest estimated risk (Table S2). Uno's IPCW approach for 5 year prediction was 0.71 [95% CI 0.69 to 0.73] at model development (apparent validation). Internal validation suggested no statistical optimism (remained 0.71 using 500 bootstrap samples), while external validation showed a slightly poorer performance (0.69 [95% CI 0.63 to 0.75], Table 3).

Time range discrimination

Harrell's concordance index (C) is commonly used to assess global performance.[24] It is calculated as a fraction where the denominator is the number of all possible pairs of patients in which one patient experiences the event first and the other later. Harrell's C quantifies the degree of concordance as the proportion of such pairs where the patient with a longer survival time has better predicted survival (lower PI). Using our time range of 0 to 5 years, Harrell's C was 0.67 [95% CI 0.66 to 0.69] at apparent validation. Again, no optimism was noted (C=0.67) and a slightly lower performance at external validation (C=0.65 [95% CI 0.62 to 0.69]). Uno's C uses a time dependent weighting that more fully adjusts for censoring (more details in appendix 2).[25] Uno's C was also 0.67 [95% CI 0.66 to 0.69] at apparent validation, 0.67 at internal validation and 0.64 [95% CI 0.60 to 0.68] for external validation in our case study.

Table 3: Performance of breast cancer model with and without PGR at 5 years in development (n=2982) and validation data (n=686)

Performance measure	Internal Validation: apparent performance		Internal Validation: performance with optimism correction by bootstrap resampling		External Validation	
	Without PGR	With PGR	Without PGR	With PGR	Without PGR	With PGR
Discrimination						
Time range						
Harrell's C (SE)	0.674 (0.660 to 0.688)	0.682 (0.668 to 0.696)	0.673	0.680	0.652 (0.619 to 0.685)	0.679 (0.648 to 0.710)
Uno's C (SE)	0.673 (0.657 to 0.689)	0.682 (0.666 to 0.698)	0.672	0.680	0.639 (0.602 to 0.676)	0.665 (0.628 to 0.702)
Fixed time						
Uno's IPCW (5 yrs)	0.712 (0.693 to 0.732)	0.720 (0.701 to 0.740)	0.710	0.717	0.693 (0.633 to 0.753)	0.722 (0.662 to 0.781)
Calibration						
Time range						
Mean calibration (O/E)	1	1	na	na	O=285; E=269.9 1.06 (0.94 to 1.19)	O=285; E=279.0 1.02 (0.91 to 1.15)
Weak calibration - Slope	Na	na	na	na	1.05 (0.80 to 1.30)	1.16 (0.93 to 1.40)
Fixed time						
Mean calibration (KM / AvgP)	1	1	na	na	KM=0.49; AvgP=0.51 1.04 (0.95 to 1.14)¶	KM=0.49; AvgP=0.50 1.02 (0.93 to 1.10)¶
Weak calibration - Slope					1.07 (0.82 to 1.32)	1.20 (0.96 to 1.44)
ICI	Na	na	na	na	0.027 (0.012 to 0.070)¶	0.021 (0.011 to 0.063)¶
E50	Na	na	na	na	0.030 (0.007 to 0.072)¶	0.007 (0.007 to 0.064)¶
E90	Na	na	na	na	0.061 (0.021 to 0.138)¶	0.072 (0.022 to 0.123)¶
Overall						
Brier	0.210 (0.204 to 0.216)¶	0.209 (0.202 to 0.215)¶	0.211	0.210	0.224 (0.210 to 0.240)¶	0.216 (0.202 to 0.232)¶
scaled Brier	14.3% (11.8% to 16.8%)¶	14.9% (12.5% to 17.7%)¶	14.0%	14.5%	10.2% (4.0% to 15.9%)¶	13.6% (7.1% to 19.1%)¶
Clinical usefulness						
Difference in model Net Benefit and treat all Net Benefit at 23% threshold	0.2674–0.2625 = 0.0049	0.2739–0.2625 = 0.0114	-	-	0.3616 – 0.3616 = 0	0.3666–0.3616 = 0.0050

na=not applicable; O=number of observed events over 5 years; E=number of expected events over 5 years; KM=Kaplan-Meier at 5 years; AvgP=average predicted risk at 5 years; ICI=integrated calibration index; E50=; E90=; ¶The 95% confidence intervals for the overall performance and calibration measures were calculated using non-parametric

bootstrap on 500 samples with replacement.

Calibration

A second important question to answer when validating a model is 'how well do observed outcomes agree with model predictions? This relates to calibration.[8, 11] Assessment of calibration is essential at external validation [3, 26]. Below we describe a hierarchy of calibration levels and its application to survival model predictions, in line with a previously proposed framework.[8]

Mean calibration

Mean calibration (or calibration-in-the-large) refers to agreement of the predicted and observed survival fraction.

Fixed time point mean calibration is typically expressed in terms of the ratio of the observed survival fraction and the average predicted risk. The observed survival fraction at the chosen time point needs to be estimated due to censoring, which can be done using the Kaplan-Meier estimator. For the external validation cohort, the Kaplan-Meier estimate of experiencing the event within 5 years was 51%, while the average predicted probability was 49%. This indicates a minor deviation from perfect mean calibration (a ratio of 1.04, 95% CI [0.95 to 1.14], Table 3).

Weak calibration

The term 'weak' refers to the limited flexibility in assessing calibration. We are essentially summarising calibration of the observed proportions of outcomes versus predicted probabilities using only two parameters i.e. a straight line. In other words, perfect weak calibration is defined as mean calibration and calibration slope of unity. Mean calibration indicates systematic underprediction or overprediction. The calibration slope indicates the overall strength of the PI, which can be interpreted as the level of overfitting (slope <1) or underfitting (slope>1).

For a fixed time point assessment of weak calibration, we can predict the outcome at 5 years for every patient but we need to determine the observed outcome at 5 years even for those who were censored before that time. One way to do this is to fit a new 'secondary' Cox model using all of the validation data with the PI from the development model as the only covariate. The calibration slope is the coefficient of the PI. In our case study it was 1.07 [95% CI 0.82 to 1.32] for the 5 year predictions, confirming very good calibration.

Moderate calibration

Moderate calibration concerns whether among patients with the same predicted risk, the observed event rate equals the predicted risk.[6] A smooth calibration plot of the observed event rates against the predicted risks is used for assessment of moderate calibration.

The relation between the outcome at a fixed time point and predictions can be visualised by plotting the predicted risk from another 'secondary' Cox model against the predicted risk from the development model.[27] The details are presented in Appendix 3 and Table S1.

The calibration plot shows good agreement between the developed and refitted models (Figure 1A). This plot can be characterized further by some calibration metrics. The Integrated Calibration Index (ICI) is the mean absolute difference between smoothed observed proportions and predicted probabilities. The E50 and E90 denote the median and the 90th percentile absolute difference between observed and predicted probabilities of the outcome.[27] For our validation cohort, we estimated ICI was 0.03 [95% CI 0.01 to 0.07], E50=0.03 [95% CI 0.007 to 0.07] and E90=0.06 [95% CI 0.02 to 0.14].

Strong calibration

Ideally, we would check for strong calibration by comparing predictions to the observed event rate for every covariate pattern observed in the validation data. However, this is hardly ever possible due to limited sample size and/or the presence of continuous predictors.

Time range calibration

Mean calibration can be assessed by comparing observed to predicted event counts, a method that is closely related to the standardized mortality ratio (SMR), common in epidemiology.[28, 29] For the validation cohort, the total number of observed recurrent free survival endpoints was 285 versus an expected number of 269.9 (ratio 1.06 [0.94 to 1.19]). This agrees with the 5-year fixed time results. For weak and moderate calibration assessment, a similar path to the fixed time approach can be followed using a Poisson model with the predicted cumulative hazard from the original Cox model as an offset. The weak calibration results gave a calibration slope of 1.05 [95% CI 0.80 to 1.30] respectively, again confirming very good calibration. Computational details are in Appendix 3.

Overall performance

Another common measure used at validation of predictions up to a fixed time point, encompassing both discrimination and calibration, is the Brier score.[30–32] This measure also involves inverse weights and is the mean squared difference between observed survival at a fixed time point (event =1 or 0) and the predicted risk by that time point.

The Brier score for a model can range from 0 for a perfect model to 0.25 for a non-informative model in a dataset with a 50% event rate by the fixed time point. When the

6

event rate is lower, the maximum score for a non-informative model is lower, which complicates interpretation. A solution is to scale the Brier score, B, at 0 – 100% by calculating a scaled Brier score as 1-B/B_0, where B_0 is the Brier score when using the same estimated risk (the overall Kaplan-Meier estimate) for all patients.[33]

At apparent validation, the Brier score was 0.210 [95% CI 0.204 to 0.216], with a null model Brier score B_0 of 0.245, so a scaled Brier score of 14.3% [95% CI 11.8% to 16.8%]. The internal validation results were very similar to the apparent validation. At external validation, the Brier score was slightly higher at 0.224 [95% CI 0.210 to 0.240] and the scaled Brier score lower at 10.2% [95% CI 4.0% to 15.9%] (Table 3).

Approaches to assess clinical usefulness

Measures of discrimination and calibration quantify a model's predictive ability from a statistical perspective. However, they fall short with regard to evaluating whether the model may actually improve clinical decision making.[34–36] Specifically, we may wish to determine whether a model is useful to support targeting of an additional treatment to high risk patients. This is what decision curve analysis aims to do by calculating the Net Benefit of a model.[36, 37] First, we need to define a clinically motivated risk threshold to decide who should be treated. For example, we may offer chemotherapy to patients with a 5-year risk of recurrence or death exceeding 20%. Using this 20% threshold, treatment benefit is obtained for patients who would die or whose cancer would recur within 5-years and have a risk ≥20%: true positive classifications. Harm of unnecessary treatment is caused to those patients who would not die or whose cancer would not recur within 5-years but have a risk ≥20%: false-positive classifications. [38]If the harm of unnecessary treatment (i.e. a false positive decision) is small then a risk threshold close to 0% is sensible, as it would lead to treating most patients. However, if overtreatment is harmful, such as major surgery, then a higher risk threshold may be apt. The odds of the risk threshold equals the harm-to-benefit ratio. Realizing this, we can now calculate the Net Benefit by calculating the proportion of true positives (that benefit) and substracting from that the proportion of false positives (that are harmed), weighted by the harm-to-benefit ratio (*w*): [38]

$$Net\ Benefit = \frac{(TP - w * FP)}{N}$$

where *TP* is the number of true-positive decisions, *FP* the number of false-positive decisions, *N* is the total number of patients and *w* is the odds of the threshold. When we are dealing with survival data, the Net Benefit can be calculated in the presence of censoring at any prediction horizon (Vickers et al, 2008).[35] For survival data *TP* and *FP* are calculated as:

$$TP(t) = [1 - S(t, X = 1)] * P(X = 1) * N$$
$$FP(t) = [S(t, X = 1)] * P(X = 1) * N$$
$$w(t) = \frac{P_t}{1 - P_t}$$

where P_t is the predicted probability at time *t* , 1 – S(t, X=1) the observed event probability for those classified as positive, and P(*X=1)* is the probability of a positive classification. Considering only one single risk threshold for evaluation of Net Benefit is usually too limited, since the perceived harms and benefits of treatment may differ between decision makers and be context-dependent. Hence, we specify a range of reasonable thresholds which would be acceptable for treatment decisions.[39] The Net Benefit can be visualised for this range of clinically relevant thresholds using a decision curve. Decision curve analysis allows us to compare the Net Benefit for different prediction models to the default strategies of treating all or no patients ('treat all' and 'treat none').[37, 40 7]

Based on previous research we focused on a range of thresholds from 14% to 23% for adjuvant chemotherapy (Figure 1B).[41] If we choose the threshold of 23% the model has a Net Benefit of 0.27. This means that the model would identify 27 patients per 100 who will have recurrent breast cancer or die within 5 years of surgery and thus require adjuvant chemotherapy. The decision curve based on the development data shows that the model Net Benefit is only marginally greater than a 'treat all' reference strategy at the highest threshold within the acceptable range of 23%. However, in the external validation dataset, the model is not useful as it has similar Net Benefit values to the 'treat all' strategy for the full range of clinically acceptable thresholds. Therefore it is unlikely that the model is useful to support decisions around adjuvant chemotherapy (Figure 1C).

All the methods we have described are summarised in the Appendix (Table S2).

Model extension with a marker

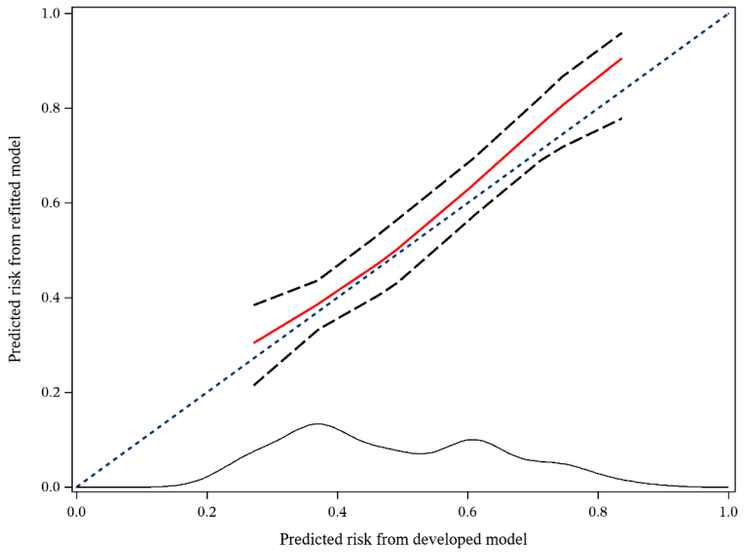
We recognize that a key interest in contemporary medical research is whether a particular marker (e.g. molecular, genetic, imaging) adds to the performance of an existing prediction model. Validation in an independent dataset is the best way to compare the performance of a model with and without a new marker. We extended our model by adding the progesterone (PGR) biomarker at primary surgery to the Cox model (Table 2). The results are described in appendix 4 and presented in Table 3. Briefly, at external validation the improvement in fixed time point discrimination was from 0.693 to 0.722 (delta AUC of 0.029), the improvement in time range discrimination was from 0.639 to 0.665 (delta C of 0.026). There was an improvement in net benefit (0.367 versus 0.362), which means we need to measure PGR in 200 patients for one additional net true positive classification.

Software

All analyses were done in SAS v 9.4 (SAS Institute Inc., Cary, NC, USA) and R version 3.6.3, R Foundation for Statistical Computing, Vienna, Austria). Code is provided at https://github.com/danielegiardiello/Prediction_performance_survival.

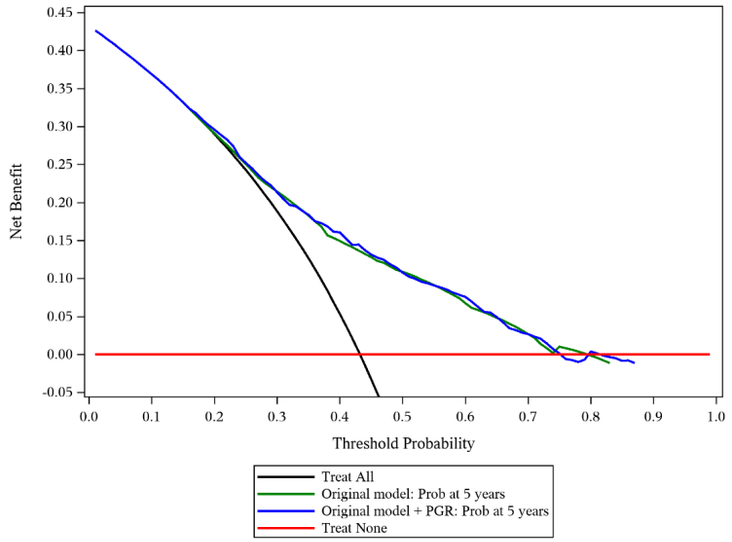
Figure 1A Calibration plot of model predicting recurrence within 5 years for patients with primary breast cancer in external validation data for fixed time assessment (A). Decision curves for predicted probabilities without (green line) and with (blue line) PGR in (B) development dataset; (C) external validation dataset.

A External validation: Fixed time assessment (predicted risk at 5 years from original model versus secondary model)

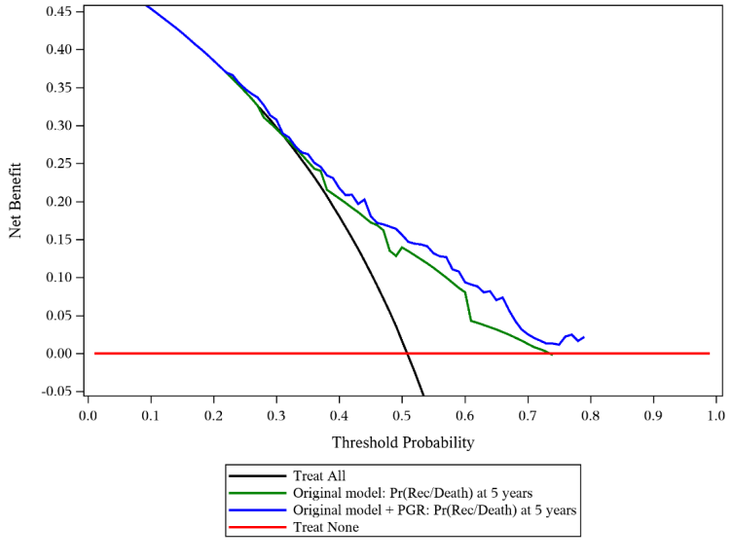


Footnote: The solid red line represents a restricted cubic spline between the predicted risk from the developed model and the predicted risk from the refitted Cox model at 5 years. The dashed lines represent the 95% confidence limits of the predicted risks from the refitted model. At the bottom of the plots is the density function for the predicted risk from the developed model.

B Decision curve analysis in development data



C Performance in external validation data



DISCUSSION

This article provides guidance for different measures that may be used to assess the performance of a Cox proportional hazards model. The performance measures were illustrated for use at model development and external validation. At model development, the apparent performance can directly be assessed for a prediction model, and internal validity is commonly assessed by cross-validation or bootstrapping techniques. External validation is considered a stronger test for a model. We first illustrated how to evaluate the quality of predictions using measures of discrimination, calibration and overall performance. We then showed how to evaluate the quality of decisions according to Net Benefit and decision curve analysis. Finally, we illustrated that the performance measures are also applicable when assessing the added value of a new predictor, where specific interest may be in improvement in discrimination and Net Benefit.

We made a distinction between measures that can be used to assess the performance of predictions for specific time points (e.g. 5- or 10-year survival) and over a range of follow up time. Prediction at specific timepoints will often be most relevant since clinicians and patients are usually interested in prognosis within a specified period of time. As described, AUC, smooth calibration curves and Brier score focus on such specific time points. Of note, estimation of the baseline survival is treated as an optional extra step in most statistical software packages. The consequence is that such key information is not available for most prediction models that are based on the Cox model. This may lead to the misconception that the Cox model does not give estimates of absolute risk. If the baseline survival for specific times points is given together with the estimated log hazard ratios, external validation is feasible (see Table S3). The discrimination and Brier score methods presented here can easily be applied to parametric survival models such as Weibull or more flexible approaches[42]

In the breast cancer study, the optimism in all performance measures was minimal at internal validation. This reflects the relatively large sample size in relation to the small number of predictors, which allows for robust statistical modeling. The performance at external validation was slightly poorer, as can in general be expected and may reflect slightly differential prognostic effects, but also differences in case-mix and censoring distribution.[43] We have not addressed the common problem of missing values for predictors, which needs somewhat more complex handling than for binary outcome prediction.[44]

Dealing with censoring is a key challenge in the assessment of performance of a prediction model for survival outcomes. If censoring is merely by end of study period ('administrative censoring'), the assumption of censoring being non-informative may be reasonable. This may not be the case for patients who are lost to follow-up, where censoring may depend on predictors in the model and other characteristics. As well as the IPCW and secondary modelling approaches presented here, other approaches are possible, for example using pseudo-observations, which often makes the assumption of fully uninformative censoring. Extensions that can deal with covariate-dependent censoring have been proposed.[45, 46]

Recommendations

We provide some recommendations for assessing the performance of a survival prediction models (Box 2 and Table S3). For calibration at external validation, we recommend plotting a smooth calibration curve (moderate calibration) and reporting both mean and weak calibration. Where no baseline survival is reported from the development study, only crude visual calibration and discrimination assessment may be possible (Appendix 5). Moreover, we recommend that researchers developing or validating a prognostic model follow the TRIPOD checklist to ensure transparent reporting.[7]

Box 2. Recommendations for assessing performance of prediction models for survival outcomes

Assessment

- For overall performance, we recommend reporting a scaled Brier score for a fixed time point assessment.
- For discrimination, report time-dependent area under the ROC curve at the time point(s) of primary interest. We recommend Uno's weighted approach. For assessment over a time range we recommend either Harrell's C or Uno's C.
- For calibration in an external dataset, while moderate calibration is essential, we recommend following the calibration hierarchy and also reporting mean and weak calibration.

Clinical utility

- If the model is to support clinical decision making, use decision curve analysis to assess the Net Benefit for a range of clinically defendable thresholds.

Publication

- When reporting development of a prediction model, include the baseline survival and ideally a link to a dataset containing the full baseline survival so others can validate the model at a fixed time point or over a range of follow up time. Report model coefficients or the hazard ratios. Both baseline survival and coefficients are essential for independent external validation of the model.
- Use the TRIPOD checklist for reporting prediction model development and validation.

Net Benefit, with visualisation in a decision curve, is a simple summary measure to quantify the potential clinical usefulness when a prediction model intends to support clinical decision-making. Discrimination and calibration are important but not sufficient for clinical usefulness. For example, the decision threshold for clinical decisions may be outside the range of predictions provided by a model, even if that model has a high discriminatory ability. Furthermore, poor calibration can ruin Net Benefit, such that using a model can lead to worse decisions than without a model.[47]

We recognize that other performance measures are available that have not been described in this paper, which may be important under specific circumstances. We recommend that future work should focus on assessing performance for various extensions of predicting survival, such as for competing risk and dynamic prediction situations.[22, 48–51]

In conclusion, the provided guidance in this paper may be important for applied researchers to know how to assess, report, and interpret discrimination, calibration and overall performance for survival prediction models. Decision curve analysis and Net Benefit provide valuable additional insight on the usefulness of such models. In line with the TRIPOD recommendations, these measures should be reported if the model is to be used to support clinical decision making.

REFERENCES

1. Hemingway H, Croft P, Perel P, et al: Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ (Online) 346, 2013
2. Meretoja TJ, Andersen KG, Bruce J, et al: Clinical prediction model and tool for assessing risk of persistent pain after breast cancer surgery. Journal of Clinical Oncology 35:1660–1667, 2017
3. Steyerberg EW, Harrell FE: Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 69:245–247, 2016
4. Altman DG, Royston P: What do we mean by validating a prognostic model? Statistics in Medicine 19:453–473, 2000
5. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Annals of Internal Medicine 130:515–524, 1999
6. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138, 2010
7. Collins GS, Reitsma JB, Altman DG, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The tripod statement. Journal of Clinical Epidemiology 68:112–121, 2015
8. van Calster B, Nieboer D, Vergouwe Y, et al: A calibration hierarchy for risk models was defined: From utopia to empirical data. Journal of Clinical Epidemiology 74:167–176, 2016
9. Sauerbrei W, Abrahamowicz M, Altman DG, et al: STRengthening analytical thinking for observational studies: the STRATOS initiative [Internet]. Statistics in medicine 33:5413–5432, 2014[cited 2021 Dec 21] Available from: https://pubmed.ncbi.nlm.nih.gov/25074480/
10. Royston P, Altman DG: External validation of a Cox prognostic model: principles and methods. Medical Research Methodology 13:33, 2013
11. Crowson CS, Atkinson EJ, Therneau TM, et al: Assessing calibration of prognostic risk scores. Statistical Methods in Medical Research 25:1692–1706, 2016
12. Rahman MS, Ambler G, Choodari-Oskooei B, et al: Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Medical Research Methodology 17, 2017
13. Stocken DD, Hassan AB, Altman DG, et al: Modelling prognostic factors in advanced pancreatic cancer. British Journal of Cancer 99, 2008
14. Foekens JA, Peters HA, Look MP, et al: The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients 1. Cancer Research 60:636–643, 2000
15. Sauerbrei W, Royston P, Look M: A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biometrical Journal 49:453–473, 2007
16. Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. Controlled Clinical Trials 17:343–346, 1996
17. Schumacher M, Bastert G, Bojar H, et al: Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. Journal of Clinical Oncology 12:2086–2093, 1994
18. Mallett S, Royston P, Dutton S, et al: Reporting methods in studies developing prognostic models in cancer:

a review. BMC Medicine 8, 2010
19. Royston P, Altman DG, Sauerbrei W: Dichotomizing continuous predictors in multiple regression: a bad idea [Internet]. Statistics in Medicine 25:127–141, 2006[cited 2021 Dec 22] Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.2331
20. Haybittle JL, Blamey RW, Elston CW, et al: A PROGNOSTIC INDEX IN PRIMARY BREAST CANCER. Br J Cancer 45:361–366, 1982
21. van Houwelingen HC: From model building to validation and back: a plea for robustness. Statistics in Medicine 33, 2014
22. Blanche P, Dartigues JF, Jacqmin-Gadda H: Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biometrical Journal 55:687–704, 2013
23. Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102:527–537, 2007
24. Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Statistics in Medicine 3:143–152, 1984
25. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30:1105–1117, 2011
26. van Calster B, McLernon DJ, van Smeden M, et al: Calibration: The Achilles heel of predictive analytics. BMC Medicine 17, 2019
27. Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 39:2714–2742, 2020
28. Breslow N, Day N: Statistical Methods in Cancer Research. Lyon, International Agency for Research on Cancer, 1987
29. Breslow NE, Lubin JH, Marek P, et al: Multiplicative Models and Cohort Analysis. Journal of the American Statistical Association 78:1–12, 1983
30. Graf E, Schmoor C, Sauerbrei W, et al: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 18:2529–2545, 1999
31. Gerds TA, Schumacher M: Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal 48:1029–1040, 2006
32. Blattenberger G, Lad F: Separating the Brier Score into Calibration and Refinement Components: A Graphical Exposition. The American Statistician 39:26–32, 1985
33. Kattan MW, Gerds TA: The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagnostic and Prognostic Research 2, 2018
34. van Calster B, Wynants L, Verbeek JFM, et al: Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology 74:796–804, 2018
35. Vickers AJ, Cronin AM, Elkin EB, et al: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Medical Informatics and Decision Making 8, 2008
36. Vickers AJ, Elkin EB: Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making 26:565–574, 2006
37. Kerr KF, Brown MD, Zhu K, et al: Assessing the clinical impact of risk prediction models with decision curves: Guidance for correct interpretation and appropriate use. Journal of Clinical Oncology 34:2534–2540, 2016
38. Peirce C: The numerical measure of success of predictions. Science 4:453–454, 1884
39. Vickers AJ, van Calster B, Steyerberg EW: Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ (Online) 352, 2016
40. Vickers AJ, van Calster B, Steyerberg EW: A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research 3, 2019
41. Karapanagiotis S, Pharoah PDP, Jackson CH, et al: Development and external validation of prediction models for 10-year survival of invasive breast cancer. Comparison with predict and cancermath. Clinical Cancer Research 24:2110–2115, 2018
42. Ng R, Kornas K, Sutradhar R, et al: The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review [Internet]. Diagnostic and Prognostic Research 2018 2:1 2:1–15, 2018[cited 2021 Dec 21] Available from: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0026-5
43. van Klaveren D, Gönen M, Steyerberg EW, et al: A new concordance measure for risk prediction models in external validation settings. Statistics in Medicine 35:4136–4152, 2016
44. Keogh RH, Morris TP: Multiple imputation in Cox regression when there are time-varying effects of covariates. Statistics in Medicine 37:3661–3678, 2018
45. Overgaard M, Parner ET, Pedersen J: Pseudo-observations under covariate-dependent censoring. Journal of Statistical Planning and Inference 202:112–122, 2019
46. Binder N, Gerds TA, Andersen PK: Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Analysis 2013 20:2 20:303–315, 2013
47. van Calster B, Vickers AJ: Calibration of Risk Prediction Models. Medical Decision Making 35:162–169, 2015
48. Bansal A, Heagerty PJ: A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. Diagnostic and Prognostic Research 3, 2019
49. Schoop R, Beyersmann J, Schumacher M, et al: Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biometrical Journal 53:88–112, 2011
50. Rizopoulos D, Molenberghs G, Lesaffre EMEH: Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 59:1261–1276, 2017
51. Wolbers M, Koller MT, Witteman JCM, et al: Prognostic Models With Competing Risks. Epidemiology 20:555–561, 2009

APPENDICES

Assessing performance in prediction models with survival outcomes: practical guidance

Appendix 1: Types of validation

Apparent performance

Apparent performance is the model's performance estimated on the same data that was used for developing the model. It is usually optimistic and therefore a poor estimate of the predictive performance in new individuals, even if those individuals are from the same population. The ultimate aim of a prediction model is to apply it on new patients for whom the outcome is still unknown. This is why it is important to conduct internal and external validation.

Internal validation

After model development it is important that we at least assess performance of the model's predictions for patients from the same underlying population.[1] The most well-known method splits the data into a model development part and a model testing part. The model is developed on the first set of data, and its performance is assessed on the second. While simple and transparent, this method is often inefficient[2]: the available data is split into two smaller parts, such that both model development and performance assessment become more uncertain. It is better to develop the model on all available data to maximize development sample size, and to use resampling methods for internal validation. The most common methods are cross-validation and bootstrapping. Cross-validation is a generalization of the split-sample method which involves splitting the data into groups. With splitting by decile, the model is estimated on 90% of the data and tested on the remaining 10%. This is repeated another 9 times, each time using the next 10% for testing. The average performance is calculated over the 10 repetitions. For more stability, such a 10-fold cross-validation procedure can be repeated 10 times (10x10-fold cv).[3] Alternatively, internal validation can be done using bootstrapping, which provides even more stable estimates of performance (at the price of increased computation time) for the population where the sample originated from. This method involves generating samples from the underlying population by drawing n samples (in the case study we used n=500) with replacement from the original dataset. Each of the n samples are the same size as the original dataset.[3] The model development process is repeated in each of the bootstrap samples and their performance assessed (bootstrap performance). Each of the models is then applied to the original dataset and test performance assessed. The average difference in the bootstrap and test performance is the 'optimism' in performance of the original model. Optimism-corrected performance is estimated as apparent performance minus optimism. It is an estimate of internal validity, reflecting validation for the underlying population where the data originated from.[4, 5]

External validation

It is preferable to have prediction models that are transportable to new (external) populations that are 'plausibly related' to those used to develop the model.[6–8] The simplest example involves the application of the model in patients from a different location. Evaluating this type of external transportability is referred to as geographical validation. Of specific interest is the evaluation of the heterogeneity in performance across many locations.[9] However, because populations at any given location tend to change over time, for example due to changes in patient care, another type of external validation involves the evaluation of a model in more recent patients from the model development location. This is referred to as temporal validation. In addition to geographical and temporal validation, it may also be relevant to determine whether a model performs well for a different type of population than the one it was developed on (domain validation).[10] For example, does a model that predicts mortality within 5 years from the point of diagnosis of early breast cancer, predict accurately for patients diagnosed with locally advanced breast cancer?.[11]

Externally validating a survival prediction model is problematic if the published article does not report the estimate of the baseline survival function for any follow-up times.

Appendix 2 Further details on methods for assessing discrimination

Time-dependent AUC

The standard approach of ROC curve analysis considers outcome status for a patient as being binary. However, in the survival setting the result depends on the timepoint of interest since the proportion of events changes over time. Recent research has incorporated this dependency on time into the estimation of sensitivity and specificity (and hence the AUC). This means that since the disease status can be observed at each time point, we may obtain different values of sensitivity and specificity throughout follow-up. This may be useful to determine how well the model performs for patients early in follow-up compared to longer term survivors. Three different approaches to estimating time-dependent sensitivity and specificity have been proposed. Each differ with regards to the time-dependent manner that the outcome status is handled.[12] In prognostic modelling the goal is generally to predict an outcome that occurs within a time period of clinical interest (e.g. within 5 years in our case study). Under this scenario we propose to focus on one suitable approach to estimate sensitivity and specificity (and hence the AUC) called 'cumulative sensitivity and dynamic specificity'. Here, at each time point each patient is classed as either a case or a non-case where a case is a patient who experiences the outcome between baseline and the time point of interest, t (e.g., 5 years), and a non-case is a patient who remains outcome free at t. The AUC evaluates whether predicted probabilities were higher for those who experience the outcome at or prior to t than for those who still have to experience the outcome.[12]

6

The Kamarudin review identified eight methods of evaluating the time-dependent AUC using the cumulative sensitivity and dynamic specificity approach and we illustrate one in our case study that is recommended by Blanche et al, 2013;[12, 13] the inverse probability of censoring weighting approach by Uno et al, 2007.[14] This approach allows us to reassign the case weights of those censored to other observations with longer follow up (see Table S1 for details of various methods for dealing with censored patients).

Concordance

Concordance (C) is one of the most popular measures of discrimination. C is defined as the fraction of all pairs of observations for which the rank order of the predictions agrees with the rank order of the actual response, i.e., the prediction model got them in the right order. Observation pairs that have the same response are not used, while pairs that have the same predicted value count as 1/2 an agreement. For a continuous response this definition is equivalent to Somers' d, for a binomial response it leads to the area under the curve (AUC), and for a survival response to Harrell's C. C is only equivalent to the AUC for binomial outcomes which has caused confusion for applied researchers who incorrectly use these terms interchangeably in the survival setting.[15] For survival data, Harrell's C is the most commonly applied, however, it does not account for censored data. Two important refinements to C for survival data are the addition of administrative censoring at the time point of interest, *t*, and the addition of a time dependent weighting that more fully adjusts for censoring.[16] If interest is focused on predicted survival up to *t*=5 years, for instance, then relative rankings between patient pairs who both have events beyond 5 years might be considered irrelevant. For the example data, the estimated 5-year concordance for prediction in the development and validation data sets was 0.674 (95% CI 0.660 to 0.688) and 0.652 (95% CI 0.619 to 0.685), respectively, using Harrell's C]). Uno's C uses a time dependent weighting that more fully adjusts for censoring. Using Uno's C, the estimated 5-year concordance was 0.673 (95% CI 0.657 to 0.689) in the development data and 0.639 (95% CI 0.602 to 0.676) in the external data. It has been shown that the bias from Harrell's C is more pronounced when it is greater than 0.8 which is rare for prediction modelling in the absence of overfitting.[17] Weighted measures such as Uno have been shown to become biased when censoring is large leading to extreme weights.[17]

Table S1: Approaches to deal with censoring in the analysis of performance at a fixed time point for a survival outcome

Approach	Concept	Assumption	Applications	Data illustration ^
Inverse probability of censoring weights (IPCW)	Set the weights of patients censored before time *t* to zero, reassigning their mass to other patients still at risk at time *t*. Can also be extended to a time dependent IPCW.	Fully uninformative censoring*	Weighted Brier score; Uno's approach to discrimination Uno's C uses a time dependent weighting (more details in appendix 2)[16]	Redistribute the weight of 280 patients who are censored before 5 years to the 406 with either an event or no event observed at 5 years
Use of a secondary model	Impute censored observations by predictions from a flexible secondary model using the complementary log-log transformed predicted risk at t years as the only covariate.	Uninformative censoring given the risk score, and proportional hazards **	Austin et al (2020) approach to calibration.[18]	Analyze 686 patients
Pseudo values	Impute censored patients by estimated survival captured in pseudo values	Fully uninformative censoring but extensions can deal with covariate-dependent censoring.	Assess calibration and discrimination with pseudo values	Analyze 686 patients (including 280 censored patients) with pseudo values

^ 280/686 GBSG (external validation dataset) subjects are censored before 5 years
* This assumption is stronger than at model development, where censoring is assumed to be uninformative given the risk score (as modeled from predictors or outcome). However, methods are available to make the weights covariate dependent[19]
** This assumption is similar to model development with Cox regression.

Table S2. Characteristics of key performance measures for the evaluation of survival prediction models

Aspect	*Fixed time point or time range*	*Measure*	*Visualization*	*Characteristics*
Discrimination	Fixed	Time-dependent (cumulative/dynamic) AUC - Uno*	Time-dependent AUC curve plots	At time, t, each patient is classed as either a *case* or a *non-case*. A case is a patient who experiences the outcome between baseline and t (or at t). A non-case is a patient who remains outcome free at t. The AUC evaluates whether predicted probabilities were higher for those who experience the outcome at or prior to t than for those who still have to experience the outcome.[14, 17, 20]
	Time range	Concordance (C) - Uno - Harrell	Kaplan Meier curves provide informal evidence of discrimination[21] (Appendix 6)	Calculated as a fraction where the denominator is the number of all possible pairs of patients in which one patient experiences the event first and the other later. C quantifies the degree of concordance as the proportion of such pairs where the patient with a longer survival time has better predicted survival.[20] Harrel's C excludes pairs where the patient with shorter follow up is censored. Uno' C adjusts more fully for censoring.[16]

Table S2 Continued

Aspect	*Fixed time point or time range*	*Measure*	*Visualization*	*Characteristics*
Calibration		*Mean calibration (calibration-in-the-large)*		Simplest type of calibration which evaluates if the observed outcome rate is equal to the average predicted risk.
	Fixed	- (1-Kaplan-Meier)/average predicted risk at t		
	Time range	- Poisson model intercept (O/E)		Use Poisson model intercept with log cumulative hazard as offset.[22]
		Weak calibration		Assesses global under or over prediction and overfitting (slope<1) or underfitting (slope>1). See appendix 3 for details on calculations.
	Fixed	- Calibration slope using secondary Cox model		
	Time range	- Calibration slope using Poisson model		Slope is coefficient of PI in Poisson model with log cumulative hazard function minus PI as offset.
		Moderate calibration		
	Fixed	- Model relationship between predictions and observed risk in external dataset using secondary Cox model - Complemented with ICI, E50, E90 - Plot of time versus O/E	Smooth calibration curve of observed t-year risk of the outcome versus predicted probability by t-years.	Reveals miscalibration which cannot be detected using calibration-in-the-large and the calibration slope approaches. Plot predicted risk of this model against predicted risk from original model.[18]
	Time range	- Model relationship between predictions and observed risk in external dataset using Poisson model	Plot the observed / expected number of events over time. Plot cumulative hazard from Poisson model versus cumulative hazard from original Cox model	Visualises O/E across all time points up to t.
Overall performance	Fixed	Brier score and scaled Brier score		Captures calibration and discrimination aspects. Interpretability is improved by scaling between 0 and 100%.
Clinical usefulness	Fixed	Net Benefit	Decision curve	Net number of true positives gained by using model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)[23]

§ PI = prognostic index; * A modified version of Uno's weighted approach is available that uses weights that are the conditional probability of being uncensored. These are calculated using the Cox model and allowing for covariate-dependent (as opposed to uninformative) censoring.[13]

Appendix 3 Calibration assessment

Calibration can be evaluated either across all follow up time points (time range assessment) or at one specific time point. Time range assessment refers to the evaluation of estimated risks at the time of the event (or censoring) for each patient. Evaluating models over the time range requires the availability of the development dataset, or at least the baseline survival for all time points. Here we describe the methods for assessment of calibration over the time range:

Mean calibration

When we wish to assess calibration across all time points then one method to deal with this is to consider a comparison of the total number of observed events (O) as compared to expected events (E), counts instead of probabilities. The expected count for each subject is defined as the predicted cumulative hazard for that subject, under the model, up until that subject's event time or censoring. This approach has a long history in epidemiology where sum(observed)/sum(expected) is known as a standardized incidence ratio (SIR).[24, 25] Such data can be analysed using standard Poisson methods and software (Berry, 1983).[26] However, in order to estimate the cumulative hazard, the dataset used to develop the original model or, at least, the baseline survival for all time points is required.[22] Failing that, linear interpolation may be used if the baseline survival is available at several time points.

For the Rotterdam dataset, there are 1275 events within 5 years of study entry. Using the German Breast Cancer Study Group (GBCSG) validation dataset, there are 285 observed events while the Rotterdam model applied to that data predicts 269.9, giving an O/E ratio of 1.06. Using the individual observed (as outcome) and expected values (log cumulative hazard as an offset term) a Poisson model, estimates an intercept term of 0.054 with a standard error of 0.059. The exponential of this value leads to exactly the same O/E estimate of 1.06, and a confidence interval of (0.94, 1.19). Fig S1A shows how O/E changes over time, remaining stable from 18 months.

Weak calibration

For binary outcomes, calibration can be inspected visually using a calibration plot of the observed proportion of outcome associated with a model's predicted risk. The PI is regressed on the observed outcomes using a logistic calibration model.[27] The coefficient of PI is the calibration slope and its value indicates whether there is overfitting (slope<1) or underfitting (slope>1).[28] Since the calibration slope does not involve grouping patients and provides a measure of the magnitude and direction of miscalibration with 95% confidence interval, it is preferred to the Hosmer-Lemeshow goodness-of-fit test, the use of which is discouraged due to focus on p-values and poor test performance characteristics.[28, 29] For the Cox model, there are different variations of the Hosmer-Lemeshow test including tests proposed by Grønnesby and Borgan,[30] which should not be used for similar reasons.

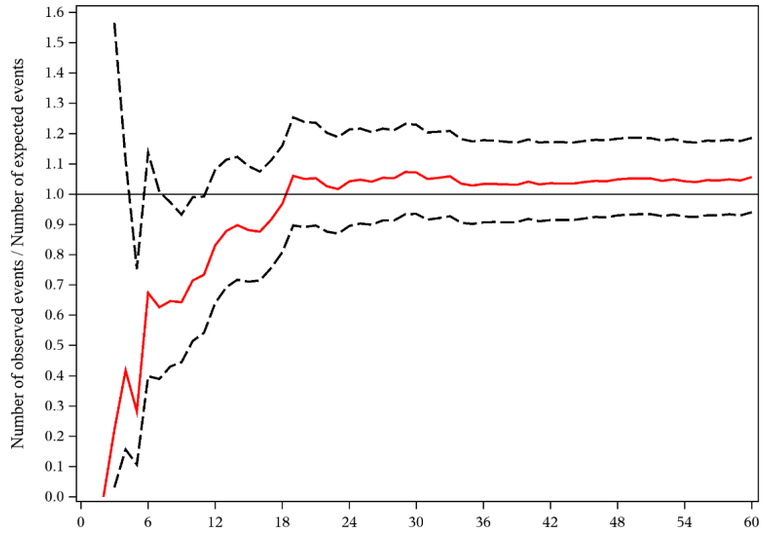


Fig S1A: Time range assessment of O/E in external dataset
Note: the solid red line represents O/E at each month up to 5 years and the dashed lines represent the 95% confidence limits of O/E

For survival outcomes, estimation of the calibration slope is possible using a Poisson model. This is done by including the PI in the validation dataset (using the coefficients from the original Cox model) as a predictor in a Poisson model with the difference between the log cumulative hazard and PI as an offset and using a log link.[22] The regression coefficient for PI represents the calibration slope. In our study the calibration slope was 1.05 (95% CI 0.80 to 1.30), so close to the ideal value of 1. The calibration intercept is just the intercept term before exponentiating in the previous section on mean calibration. This approach is termed weak calibration because of its limited flexibility in assessing calibration. We are essentially summarising calibration (of the observed proportions of outcomes versus predicted probabilities) using only two parameters. However, more subtle violations of miscalibration may remain undetected.

Moderate calibration

The relation between the outcome over the time range and predictions can be visualised by plotting the predicted cumulative hazard from the Poisson model against the predicted risk from the development model. In the external dataset, the PI from

the original Cox model is modelled as a restricted cubic spline in a Poisson model with the log of the cumulative hazard as the offset. Predictions from this Poisson model represent a proxy to the observed outcomes for all patients including those who were censored. The calibration plot shows good agreement between the Cox and Poisson models (Fig S1B).

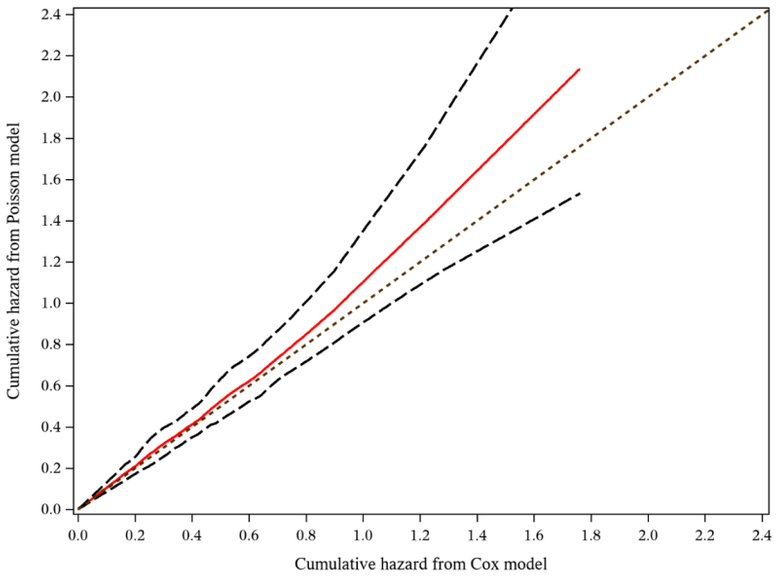


Figure S1B: Calibration plot of predicted cumulative hazard of recurrence over the time range for Cox model versus Poisson model
Note: the solid red line represents the relationship between the predicted cumulative hazard from the developed model and the predicted cumulative hazard from the Poisson model. The dashed lines represent the 95% confidence limits of the predicted cumulative hazard from the Poisson model.

Appendix 4 Incremental value of PGR

We extended the model by adding the progesterone (PGR) biomarker at primary surgery to the Cox model. Following examination for non-linearity, PGR was fitted as a restricted cubic spline function with 3 knots (see Figure S2). We repeated the apparent, internal and external validation processes on this extended model.

Performance in development dataset

PGR had additional predictive value when added to the original model, increasing the model chi-squared from 483.7 to 516.7 (LR statistic 33.0, df=2, P<0.001) in the development dataset. Overall performance showed a small increase: Brier score decreased from 0.210 to 0.209, and the scaled Brier score increased from 14.3% to 14.9% (Table 4). The discriminative ability at 5 years follow-up also increased marginally (e.g., Uno's weight approach increased from 0.712 to 0.720).

For a threshold of 23%, the model with PGR included had a slightly larger net benefit than the model without PGR (0.274 versus 0.267) (Figure 1B). Hence, at this particular cut-off, the model with PGR would be expected to lead to one more net true positive classification per 154 patients (1/0.0065) at the same number of false positive classifications.

Performance in external dataset

Comparing the above performance measures for the model with and without PGR in the external dataset, the former was better overall. The improvement in fixed time point discrimination was from 0.693 to 0.722 (delta AUC of 0.029) at external validation while improvement across the time range was from 0.639 to 0.665 (delta C of 0.026). Globally, the total number of observed recurrent free survival endpoints was 285 versus an expected number of 279.0. Using the Poisson model this equated to a calibration-in-the-large SIR of 1.02 (95% CI 0.91 to 1.15)). The calibration slope was 1.16 (95% CI 0.93 to 1.40). Mean calibration on average showed some improvement with PGR included. The calibration plot of O/E across all time points up to 5 years shows relatively consistent results from 18 months onwards (Figure S3A). The calibration plot of the predicted cumulative hazard in the original Cox model versus the Poisson model shows good agreement, although some underprediction in the higher risk patients (Figure S3B). Focusing on calibration at the fixed time point of 5 years we found that the Kaplan-Meier estimate of experiencing the event within 5 years was 0.49, while the average predicted probability was 0.50. The calibration plot (Figure S3C) shows evidence of good agreement overall for predictions of mortality over 5 years. The ICI decreased from 0.03 to 0.02 when PGR was included and E50 dropped from 0.03 to 0.01. The scaled Brier score increased from 10.2% to 13.6% at external validation. Hence a substantial improvement in statistical performance was found.

With PGR in the model, the risk groups are well separated in both the development and validation datasets which implies that the model discriminates well in these cohorts (Figure S4). However, from approximately 3 years into follow-up the middle two risk groups converge for the external dataset.

In the external dataset, the net benefit was similar for models with or without PGR (Figure 1C). However, at the risk threshold of 23% the model without PGR was no better than treating all patients. The model with PGR had a slightly larger net benefit (0.367 versus 0.362), or one additional net true positive classification per 200 patients (1/0.005) at the same number of false positive classifications.

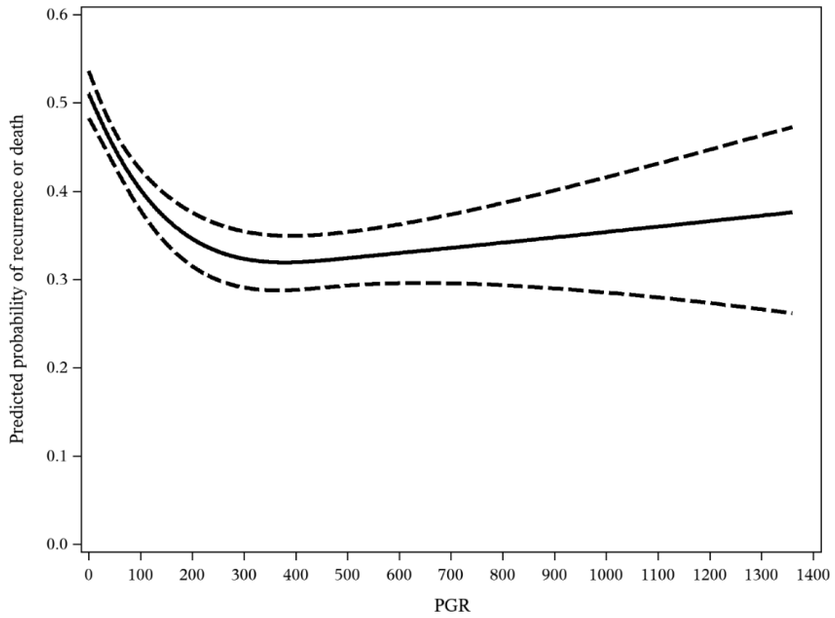
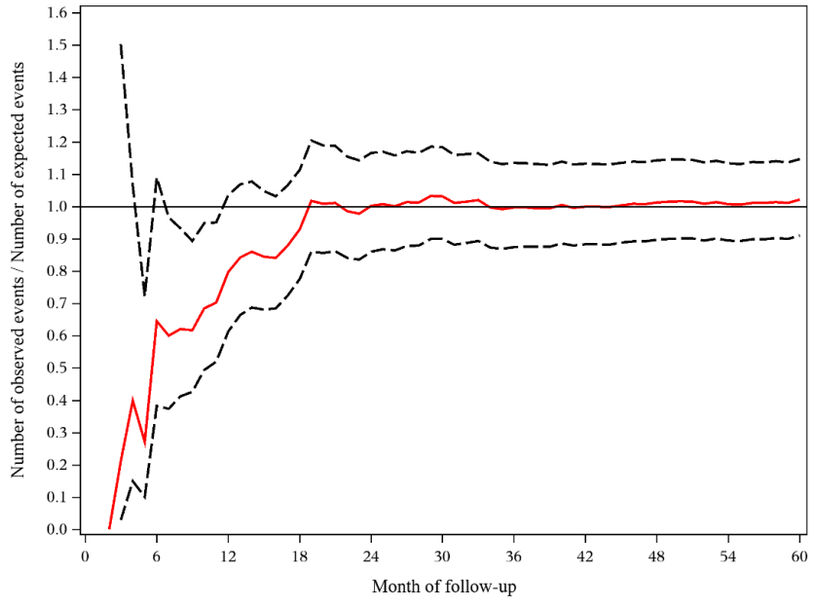
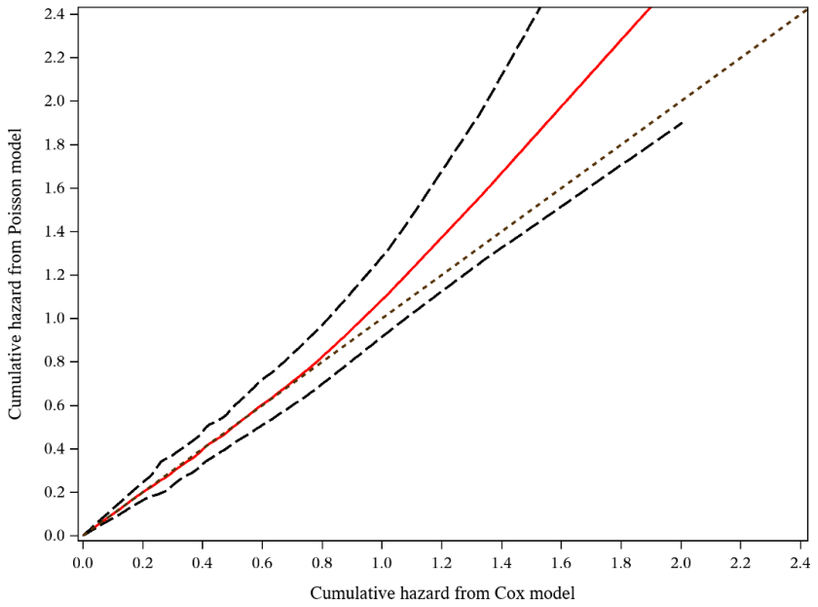


Figure S2: Plot showing unadjusted (univariable) relations between PGR and predicted probability of recurrence (solid curve) with 95% confidence bands. The relation was non-linear characterised by a restricted cubic spline function with 3 knots.

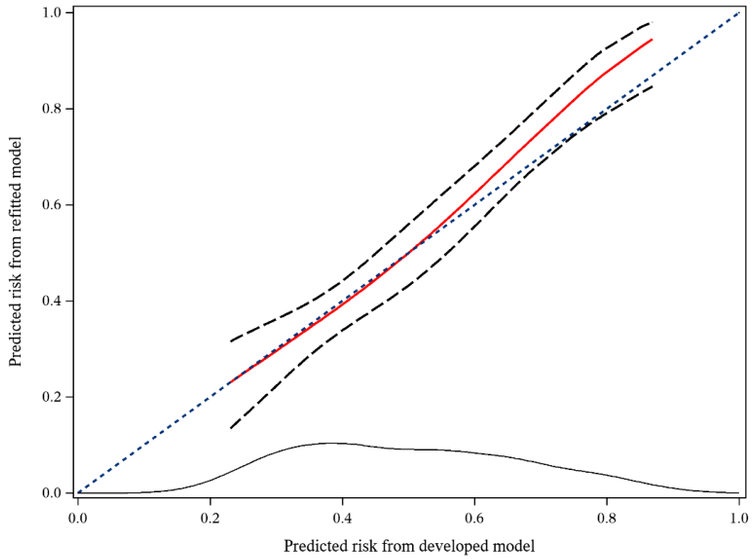
Figure S3: Calibration plots of Cox model with PGR predicting recurrence within 5 years for patients with primary breast cancer in the external validation data.



A O/E across the time range



B Predicted cumulative hazard from original model versus Poisson model



C Predicted risk from original model versus secondary model at 5 years

Note: In A, the solid red line represents O/E at each month up to 5 years and the dashed lines represent the 95% confidence limits of O/E; In B, the solid red line represents the relationship between the predicted cumulative hazard from the developed model and the predicted cumulative hazard from the Poisson model. The dashed lines represent the 95% confidence limits of the predicted cumulative hazard from the Poisson model. In C, the solid red line represents a restricted cubic spline between the predicted risk from the developed model and the predicted risk from the refitted model at 5 years. The dashed lines represent the 95% confidence limits of the predicted risks from the refitted model. At the bottom of the plots is the density function for the predicted risk from the developed model.

Table S3 What calibration assessments can I do based on the model development information I have?

What development data do you have?	Fixed timepoint assessment	Continuous time assessment	Methods
Whole dataset used to develop model	✓	✓	See section on calibration and appendix 3 for calibration methods
Table of baseline survival at all observed time points + PI	✓	✓	See section on calibration and appendix 3 for calibration methods
Baseline survival at multiple (but not all) time points (e.g., yearly) + PI	✓	✓	Use interpolation methods to estimate baseline survival (Crowson et al, 2016).[22] Then see section on calibration and appendix 3 for calibration methods.
A predicted survival curve based on the model + PI	✓	✓	Use digitisation software to estimate baseline survival (Guyot et al, 2012).[31] Then see the section of calibration and appendix 3 for calibration methods.
Baseline survival at time point of interest + PI	✓		See section on calibration and appendix 3 for calibration methods at fixed time points.
Published Kaplan-Meier curves for risk groups			Formal assessment not possible. Can visually compare Kaplan-Meier curves to those from validation data (Appendix 5; Royston and Altman, 2013)[21]
None of the above			Calibration assessment not possible

Appendix 5 What to do if the development dataset (or its baseline hazard) is not available

In case the baseline hazard/survival function (either as a look-up table or mathematical function) of a survival model is not available then there is not enough information to formally assess calibration. However, if the development paper reported Kaplan-Meier curves for risk groups of the PI then it is possible to compare these with the corresponding Kaplan-Meier curves from the validation cohort.[21, 32] This is not a strict comparison between observed and predicted values since we are using Kaplan-Meier estimates and not the Cox model-based predictions. If the survival curves for risk groups overlap between the development and validation datasets, then this may provide an indication of agreement. Further, plots where the curves are widely separated between risk groups provides informal evidence of discrimination.

In the case study, we centred the PI for the model including PGR at average risk by subtracting its mean of 0.65 and then categorised it into quarters. The groups at the extreme ends represent the lower and upper fourth of the risk of recurrence. This procedure was done in both the development and validation datasets. For the validation dataset the PI was calculated based on the coefficients from the model fitted to the development dataset (Figure S4). In the development dataset the four risk groups are well separated which implies that the model has discriminative ability in this cohort. However, the curves for the second and third fourths are close together in the validation data suggesting that the model does not discriminate well between these two groups. Otherwise, the discrimination is broadly similar between the two datasets. The curves do not agree too well in absolute risks between the two datasets suggesting that there is a degree of miscalibration. The percentage of patients within the four groups in the validation dataset were 8.8%, 21.0%, 36.3% and 34.0% respectively so there are more in the two highest risk fourths and fewer in the lowest risk fourth than in the development dataset. The mean (SD) PI was 0.24 (0.50) in the validation dataset, implying that the prognostic profile was somewhat worse than in the development dataset. This is evident from Table 1 which shows that women in the validation dataset had larger tumours and more nodes.

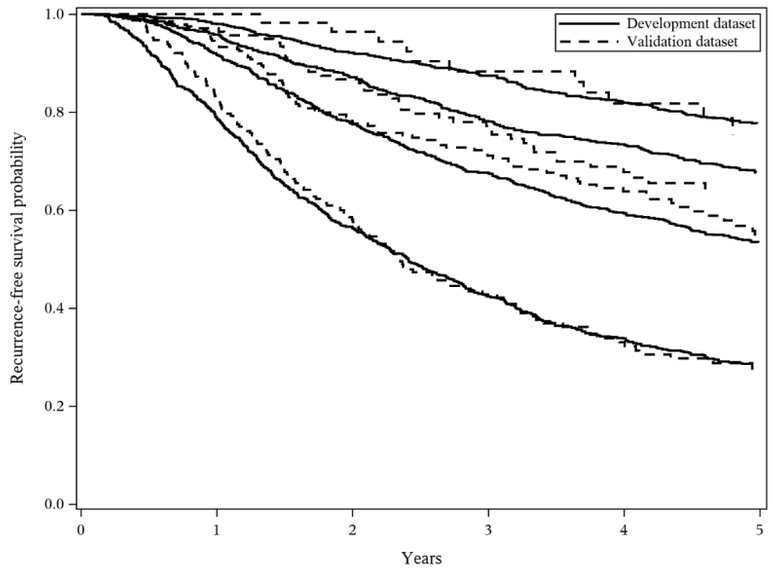


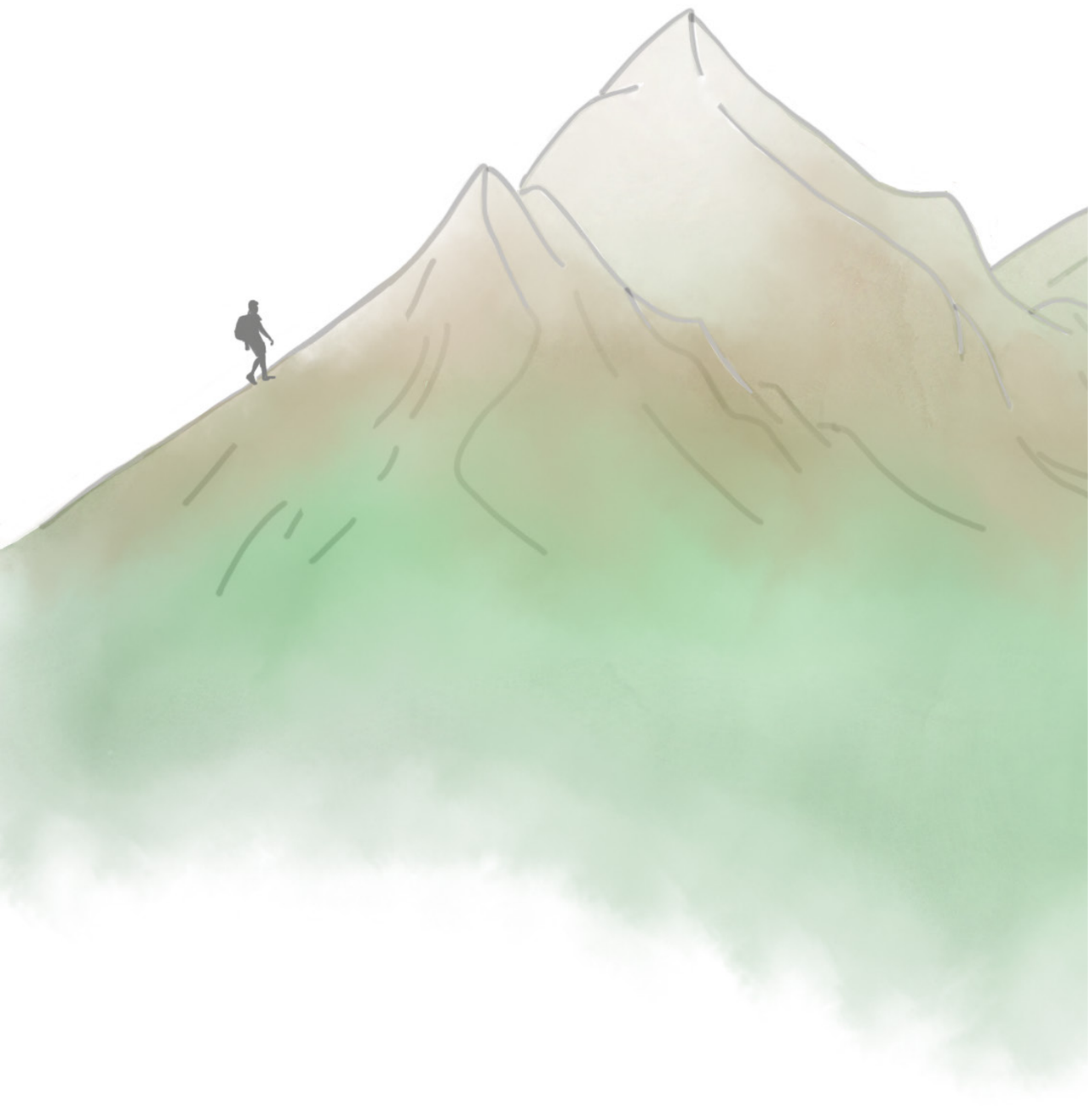
Fig S4 Kaplan-Meier curves for event-free survival in 4 equal sized risk groups in the development and validation cohorts for model with PGR

REFERENCES

1. Altman DG, Royston P: What do we mean by validating a prognostic model? Statistics in Medicine 19:453–473, 2000
2. Steyerberg EW, Harrell FE: Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 69:245–247, 2016
3. Harrell FE: Regression Modeling Strategies [Internet]. Cham, Springer International Publishing, 2015[cited 2021 Jun 9] Available from: http://link.springer.com/10.1007/978-3-319-19425-7
4. Efron B, Tibshirani RJ: An Introduction to the Bootstrap. Boston, MA, Springer US, 1993
5. Harrell FE, Lee KL, Mark DB: TUTORIAL IN BIOSTATISTICS MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. 1996
6. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Annals of Internal Medicine 130:515–524, 1999
7. Austin PC, van Klaveren D, Vergouwe Y, et al: Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. Journal of Clinical Epidemiology 79, 2016
8. Siontis GCM, Tzoulaki I, Castaldi PJ, et al: External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology 68, 2015
9. Steyerberg EW, Nieboer D, Debray TPA, et al: Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Statistics in Medicine 38, 2019
10. Toll DB, Janssen KJM, Vergouwe Y, et al: Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology 61, 2008
11. Gray E, Donten A, Payne K, et al: Survival estimates stratified by the Nottingham Prognostic Index for early breast cancer: A systematic review and meta-analysis of observational studies 11 Medical and Health Sciences 1117 Public Health and Health Services 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis. Systematic Reviews 7, 2018
12. Kamarudin AN, Cox T, Kolamunnage-Dona R: Time-dependent ROC curve analysis in medical research: Current methods and applications. BMC Medical Research Methodology 17, 2017
13. Blanche P, Dartigues JF, Jacqmin-Gadda H: Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biometrical Journal 55:687–704, 2013
14. Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102:527–537, 2007
15. Blanche P, Kattan MW, Gerds TA: The c-index is not proper for the evaluation of t-year predicted risks. Biostatistics 20:347–357, 2019
16. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30:1105–1117, 2011
17. Schmid M, Potapov S: A comparison of estimators to evaluate the discriminatory power of time-to-event models. Statistics in Medicine 31:2588–2609, 2012
18. Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 39:2714–2742, 2020

19. Gerds TA, Kattan MW, Schumacher M, et al: Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Statistics in Medicine 32:2173–2184, 2013
20. Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Statistics in Medicine 3:143–152, 1984
21. Royston P, Altman DG: External validation of a Cox prognostic model: principles and methods. Medical Research Methodology 13:33, 2013
22. Crowson CS, Atkinson EJ, Therneau TM, et al: Assessing calibration of prognostic risk scores. Statistical Methods in Medical Research 25:1692–1706, 2016
23. Vickers AJ, Cronin AM, Elkin EB, et al: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Medical Informatics and Decision Making 8, 2008
24. Breslow N, Day N: Statistical Methods in Cancer Research. Lyon, International Agency for Research on Cancer, 1987
25. Breslow NE, Lubin JH, Marek P, et al: Multiplicative Models and Cohort Analysis. Journal of the American Statistical Association 78:1–12, 1983
26. Berry G: The analysis of mortality by subject-years method. Biometrics 39:173–184, 1983
27. Cox D: Two further applications of a model for binary regression. Biometrika 45, 1958
28. van Calster B, Nieboer D, Vergouwe Y, et al: A calibration hierarchy for risk models was defined: From utopia to empirical data. Journal of Clinical Epidemiology 74:167–176, 2016
29. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138, 2010
30. Grønnesby JK, Borgan Ørnulf: A method for checking regression models in survival analysis based on the risk score. Lifetime Data Analysis 2, 1996
31. Guyot P, Ades A, Ouwens MJ, et al: Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology 12, 2012
32. van Houwelingen HC: Validation, calibration, revision and combination of prognostic survival models ‡. 2000

6



Chapter 7

Validation of prediction models in presence of competing risks: a guide through modern methods

British Medical Journal. 2022 May; 24; 377:e069249
https://www.bmj.com/content/377/bmj-2021-069249

Nan van Geloven
Daniele Giardiello
Edouard F Bonneville
Lucy Teece
Chava L Ramspek
Maarten van Smeden
Kym IE Snell
Ben van Calster
Maja Pohar-Perme
Richard D Riley
Hein Putter
Ewout W Steyerberg
on behalf of the STRATOS initiative

Glossary	
Patients	Where we refer to 'patients' one can also read individuals, participants or subjects. We use 'patients' to match our illustration using breast cancer data.
Competing risks	In the competing risks setting there are multiple event types that 'compete' for first occurrence. In the case study these are breast cancer recurrence and mortality before recurrence.
Primary event	We assume one event type is the primary event of interest. In the case study, the primary event is breast cancer recurrence.
Prediction horizon	The specified duration of time over which predictions are made. In the case study we focus on 5-year risks.
Cumulative incidence	The absolute risk of experiencing the primary event during the prediction horizon, taking into account that a patient who experiences a competing event will never experience the primary event.
Primary event indicator	A patient's primary event status by the end of the prediction horizon: did the patient experience the primary event before or at that time-point? If so, the primary event indicator is 1. If the event indicator is 0, this may mean either that the patient has not experienced any event by the end of the prediction horizon or that the patient experienced a competing event by that time point.
Censoring	When the patient's event status by the end of the prediction horizon is unknown, e.g. due to loss to follow up at an earlier time point.
Observed outcome proportion	This is the observed proportion of patients with the primary event. In a setting without censoring, this is the sum of the primary event indicators divided by the total number of patients. With censoring, the observed outcome proportions have to be estimated accounting for the incomplete observations. The observed outcome proportion represents the actual underlying cumulative incidence.
Risk estimates (or estimated risks)	These are the estimates of cumulative incidence from the developed prediction model. Typically, risks up to one or a few time-points are of particular interest. We want to evaluate the performance of these risk estimates for new patients.

Stand first

Thorough validation is pivotal for any prediction model before it is advocated for use in medical practice. For time-to-event outcomes such as breast cancer recurrence, death from other causes is a competing risk. Model performance measures must account for such competing events. In this paper, we present a comprehensive yet accessible overview of performance measures for this competing event setting, including the calculation and interpretation of statistical measures for calibration, discrimination, overall prediction error, and clinical utility by decision curve analysis. All methods are illustrated for patients with breast cancer, with publicly available data and R code.

Key messages

- Validation is a necessary step for prediction models before being used in clinical practice.
- In the presence of competing risks, these other risks have to be accounted for at model validation.
- We provide a comprehensive overview of performance measures for calibration, discrimination, overall prediction error and decision curve analysis that account for competing events.
- Data and R code used for illustration of the measures is available from a GitHub page.

INTRODUCTION

Prediction models are pivotal for counseling patients about their prognosis and for risk stratification.[1] Interest often lies in predicting a non-fatal adverse event over a certain time period, e.g. breast cancer recurrence within 5 years after diagnosis. As study populations of common diseases increasingly consist of elderly individuals with high degrees of multimorbidity, patients will experience other events that preclude the occurrence of the event of interest.[2] For example, a patient with a previous breast cancer who dies from a cardiovascular cause can no longer experience breast cancer recurrence. In these settings prediction models should target the *cumulative incidence* (or absolute risk[3]) of the adverse event, which is defined as the probability of the event of interest occurring by a particular time-point with no other competing event occurring earlier. In the breast cancer example, the 5-year cumulative incidence of recurrence is the risk of developing a recurrence within 5 years taking into account that patients who die within 5 years cannot develop recurrence anymore. Failing to account for competing events during model development leads to overestimation of the cumulative incidence.[4] The higher the risk of the competing event, the more pronounced the overestimation. Crucially, failing to account for competing events during validation leads to a distorted view on model performance, especially for calibration. This was recently revealed for an internationally recommended kidney failure prediction model, which systematically overestimated 5-year absolute risk of kidney failure in patients with advanced chronic kidney disease. The absolute overestimation by 10 percentage points on average and by 37 percentage points in the highest risk group could result in overtreatment of patients and therefore led to the conclusion that the model was unfit for use in this population. This was missed in previous validation efforts which ignored the competing event of death.[5,6] We present model performance obtained when ignoring the competing risk and when accounting for it side-by-side in Supplementary material 1.

For predicting binary and time-to-event outcomes, useful guidance on how to perform model validation exists.[7-10] For time-to-event outcomes with competing risks, validation guidance is currently spread out over many technical papers which hampers the uptake of appropriate methods in medical research. We aim to provide an accessible overview of contemporary performance measures for time-to-event outcomes with competing risks. Our overview was made on behalf of the international STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (http://stratos-initiative.org), which aims to provide guidance documents for relevant topics in the design and analysis of observational studies for a non-specialist audience.[11] We focus on how to calculate and interpret performance measures with illustration using a breast cancer prediction model, including accompanying R code.

SETTING

In this paper, we assume a prediction model has already been developed. It should have been reported such that it allows calculating the estimates of the cumulative incidence (or absolute risk of an event) at the time point(s) of interest for new patients (Supplementary material 2). Our aim is to validate this model in an external dataset while accounting for competing events. Our focus is on external validation studies. The same performance measures could also be used during internal validation when combined with techniques such as bootstrapping or cross-validation.[12] Typically, interest is in the evaluation of the prediction of the primary event occurring by a single specific time-point. If multiple time-points are of interest clinically, we may assess performance at each of these time-points or over a time range until the last time-point of interest.

Breast cancer case study

For illustration, we consider a simple competing risks prediction model for the cumulative incidence of breast cancer recurrence within 5 years after diagnosis developed on the FOCUS cohort, a Dutch cohort of consecutive breast cancer patients aged 65 years and older. We used cause-specific Cox proportional hazards regression modeling with the following four predictors: age at diagnosis, tumor size, nodal status, and hormone receptor status (Supplementary material 2 and Table 1).

Table 1: Hazard ratios for the developed prediction model

Predictor at breast cancer diagnosis	Cause-specific hazards models	
	Recurrence	Other cause mortality
	cHR (95%CI)	cHR (95% CI)
Age, years (80 vs 69)	1.18 (0.90-1.55)	3.41 (2.76-4.24)
Size, cm (3.0 vs 1.4)	1.49 (1.25-1.78)	1.46 (1.26-1.70)
Nodal status (positive vs negative)	1.66 (1.18-2.35)	1.20 (0.91-1.60)
HR status (ER-/PR- vs ER and/or PR+)	1.90 (1.31-2.78)	1.27 (0.90-1.80)
5 year baseline cumulative incidence	0.14	0.18

Abbreviations: cHR: cause specific hazard ratio; CI: confidence interval; HR: hormone receptor; ER: estrogen receptor status; PR: progesterone receptor status. For representation purposes, the cHR for the continuous predictors (age and size) are listed for the 75th versus the 25th percentile. Baseline cumulative incidence is presented at the overall mean of the linear predictor in the model. To estimate the 5-year cumulative incidence of recurrence for a new patient, first for each event the patient's predictor values are multiplied by the cause-specific (log) hazard ratios and combined with the cause-specific baseline hazards. Secondly, the resulting cause specific hazards for both events are combined over time up to and including 5 years (Supplementary materials 2 and 4).

We assess the performance of this model in patient data from the Netherlands Cancer Registry (NCR), which is a different dataset to that used for model development. We selected patients aged 70 years or older diagnosed with breast cancer between 2003 and 2009 in the Netherlands who received primary breast surgery, and received no previous neoadjuvant treatment. We used a random subset of 1000 patients from the registry as with this selection we were allowed to share the individual patient data open access. Among these 1,000 patients, 103 recurrences and 187 non-recurrence deaths occurred within 5 years follow up (cumulative incidence curve in Supplementary Figure 1).

Performance measures for risk prediction models and accounting for competing risks

We discuss performance measures for the following four validation aspects: calibration, discrimination, overall prediction error and decision curve analysis. We give the results of these performance measures in our breast cancer case study. Corresponding R functions are in Table 2, and technical descriptions in Supplementary material 4.

Calibration

Calibration refers to the agreement between observed outcome proportions and risk estimates from the prediction model. For example, in the breast cancer cohort, the model predicted a 14% absolute risk of breast cancer recurrence by 5 years on average. This implies that, if the model is well calibrated on average, we expect to observe a recurrence event in approximately 14% of the patients in the validation set within 5 years. Ideally calibration is not only adequate on average ('calibration in the large'), but across the entire range of predictions.

Calibration plot

Calibration plots offer a detailed view on calibration by comparing observed and predicted outcomes among patients with the same estimated risk. The observed outcome proportions and estimated risks by a particular time-point of interest are plotted against each other, with deviations from the diagonal signalling miscalibration. A common approach is one where individuals are divided into approximately equally sized groups based on their risk estimates - for example in tenths of risk defined between deciles. Then, for each group, the observed outcome proportion is plotted against the estimated risk. The main challenge is how to incorporate censored data and competing events into the calculation of the observed outcome proportion. When using the grouping approach, the observed outcome proportion can be estimated using the Aalen-Johansen estimator (Supplementary material 4).[13-15] However, as the grouping approach has been criticized for arbitrariness of the categorization and potential loss of information, it is recommended to include a smoothed curve in the calibration plot.[16] One approach of obtaining a smooth curve is using pseudo-observations. These pseudo-observations replace the primary event indicators, which gives a proxy 'observed' event indicator for all patients, even those that were censored observations (Box 1).[17] After this transformation into pseudo-observations, a smooth curve can be obtained using a non-parametric smoother of the observed outcome proportions (from the validation

Table 2: Overview of performance measures with suggested R packages that offer implementation for competing risk outcomes

Aspect	Performance measure	Interpretation		R package (function)
Calibration	calibration plot	How close is each estimated risk (or risk group) to the observed outcome proportion?		riskRegression (plotCalibration)
	O/E ratio	How close is the estimated risk to the overall observed outcome proportion?	Calibration in the large ('mean calibration'): ratio of average estimated risk to overall observed outcome proportion	available from our GitHub
	calibration intercept		Intercept (on the log cumulative-hazard scale) of the regression of observed outcomes with estimated risks as offset	
	calibration slope	Are estimated risks too extreme (far apart) or too modest (homogeneous)?	Slope (on the log cumulative-hazard scale) of the regression of observed outcomes on estimated risks	
Discrimination	c-index	How well does the model separate those who experience the primary event earlier than others?		pec (cindex)
	C/D AUC_t	Cumulative/dynamic area under the receiving operator characteristic curve. How well does the model separate those who will and who will not experience the primary event by a certain time-point?		timeROC (timeROC)
	C/D AUC_t curve	C/D AUC_t calculated for each time-point up to the time-point of interest		available from our GitHub
Prediction error	Brier score	How close are estimated risks to the observed primary event indicators?	Average squared difference between estimated risks and primary event indicators	riskRegression (Score)
	scaled Brier score		Percentage reduction in Brier score compared to a null model	
Decision curve analysis	Net Benefit	What is the net result from correctly and falsely classified high risk patients?	Weighted difference between correctly and falsely classified patients, for a certain risk threshold	available from our GitHub
	Decision curve		Curve of Net Benefit over a plausible range of risk thresholds	

data) versus estimated risks (from the model).[18;19] An alternative approach was recently proposed where the smoothed curve is obtained as predictions from a flexible regression model (Box 1).[20,21] Both for the pseudo-observations approach and for the flexible regression approach, the calibration curve will depend on the chosen strength of the smoothing, i.e. the span for the first approach and the degree of flexibility (e.g. number of knots when using splines) in the second approach. Advice on these choices can be found elsewhere.[18,21] The smoothed curve should only be plotted over the range of observed risks and not extrapolated beyond.

The calibration plot for the breast cancer model shows that the predicted 5-year cumulative incidence of breast cancer recurrence is too high at the lower range of the estimated risks in the validation cohort (Figure 1, estimated using the pseudo-observations approach). The calibration curve using the flexible regression approach showed similar overestimation (available from our GitHub page).

Box 1: Techniques for estimating performance measures from competing risks data in the presence of censoring.

Pseudo-observations

- A pseudo-observation is used as a proxy measure of the primary event indicator of each patient.
- The pseudo-observations are calculated as the weighted difference between the cumulative incidence estimate at the prediction horizon based on all patients and the same quantity estimated leaving that patient out.
- The advantage of pseudo-observations is that censored patients for who the primary event indicator is unknown, will have a pseudo-observation and can contribute to the calculation of the observed outcome proportion in a straightforward way.

Smoothing using a flexible regression model

- The primary event is regressed on (a complementary log-log transformation of) the risk estimates, employing restricted cubic splines to allow a non-linear relationship. The shape and degree of smoothing is chosen by specifying the number and location of knots. Austin et al. propose to use a Fine and Gray model in this step.[20,21]
- Observed outcome proportions are estimated using the flexible regression model for all patients, including patients with a censored event status.

Inverse probability of censoring weighting (IPCW)

- The intention with IPCW is to create a hypothetical population that would have been observed had no censoring occurred.
- This can be achieved by up-weighting patients who are similar to censored patients but remain in the study longer. In other words, observations that were not likely to remain in follow-up are up-weighted.
- The weights are estimated from a survival model with censoring as the outcome.
- Observations are then weighted inversely to their probability of not being censored.

Numerical summaries of calibration

A simple way to summarize overall calibration (or calibration-in-the-large) by a particular time-point, is a ratio of observed and expected outcomes (O/E ratio). An O/E of 1 indicates perfect calibration-in-the-large, an O/E < 1 indicates that on average the model predictions are too high, and an O/E > 1 indicates that on average the model predictions are too low. In the presence of competing events, the O/E ratio can be calculated as the

ratio of the observed outcome proportion by the prediction horizon (estimated by the Aalen-Johansen estimator[13]) and the average risk estimated by the prediction model under evaluation. We refer to Supplementary material 3 for an overview of alternative ways to summarize overall calibration.

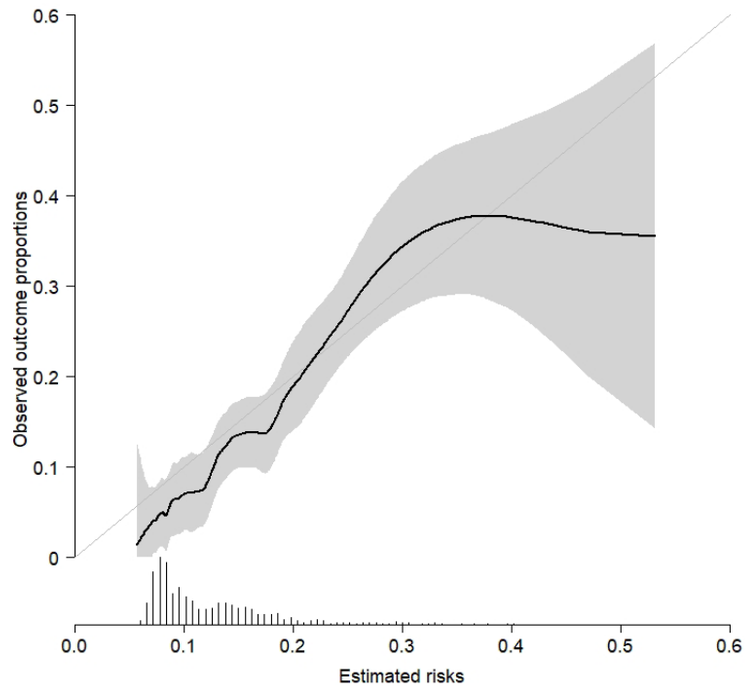


Figure 1: Calibration plot visualizing the estimates of cumulative incidence of breast cancer recurrence against the outcome proportions observed in the validation set. The 45 degree reference line indicates perfect calibration. The smooth curve was estimated using a linear loess smoother on the pseudo-observations with span of 0.33. The open dots along the x-axis indicate the distribution of risk estimates.

A further way to numerically summarize the calibration plot of predictions by a particular time-point is by calculating the calibration intercept and calibration slope. For competing risks data, these can be estimated using pseudo-observations, similar to those proposed for ordinary survival.[19] We provide details in Supplementary material 3. If on average the risk estimates equal the observed outcome proportions, the calibration intercept will be zero. The calibration slope equals one if the strength of the predictors match the observed strength in the validation set. The calibration intercept and slope can potentially be used for recalibration of existing models to fit better in new populations.[22,23]

Returning to the breast cancer validation cohort where we focus on the cumulative incidence of recurrence up to 5 years, we observe a somewhat too high estimated risk on average with an O/E ratio of 0.81 [95% CI 0.62 to 0.99]. The calibration intercept was estimated at -0.15 confirming the overestimation. For example, for an estimated risk of 14%, the observed outcome proportion was 1-0.86^(exp(-0.15))=12%. The calibration slope was 1.22 [95% CI 0.84 to 1.60], which would indicate slightly too homogeneous predictions but the wide confidence interval precludes any firm conclusions.

Table 3: Estimated values (95% confidence interval) of the performance measures in the external breast cancer data. O/E ratio: ratio of observed and expected outcomes, C/D AUCt: cumulative/dynamic area under the receiving operator characteristic curve

Calibration	O/E ratio	0.81 (0.62 to 0.99)
	Calibration intercept	-0.15 (-0.36 to 0.05)
	Calibration slope	1.22 (0.84 to 1.60)
Discrimination	c-index up to 5 years	0.71 (0.67 to 0.76)
	C/D AUCt at 5 years	0.71 (0.66 to 0.77)
Prediction error	Brier score	0.09 (0.04 to 0.13)
	Scaled Brier score	5.7% (1.6% to 8.2%)
Decision curve analysis	Net Benefit at 20% threshold	0.014

Discrimination: C-index and area under the ROC curve

As well as being well calibrated, useful prediction models should assign higher risk estimates to patients who will experience the primary event earlier than others. This is their discriminative ability.

A commonly used performance measure for assessing discrimination over a certain time range is the c-index, also known as concordance index. The c-index assesses the ordering of predictions for all patient pairs where at least one has the event within the prediction horizon and the other is not censored earlier than that event.[24] The c-index is the proportion of these examinable pairs for which the patient with the highest estimated risk is observed to experience the event sooner than the other patient. Other versions of the c-index have been proposed that are less dependent of the study specific censoring mechanism.[25,26] The c-index ranges from 0.5 (no discriminating ability) to 1.0 (perfect ability to discriminate between patients with different outcomes).

In the competing risks setting, two definitions of comparison pairs have been considered (Supplementary material 4).[27] When the target is evaluating cumulative incidence, we propose to compare pairs where one individual has the primary event within the

prediction horizon and the other either has the primary event later or experiences a competing event. Such a pair is considered concordant when the first individual has the higher estimated risk. In the presence of censoring, inverse probability of censoring weighting (IPCW) methods can be applied for estimating the c-index (Box 1).[28,27]

If interest is not in the full range of observed follow up but only in the ability of a model to predict the event occurring by a single time-point of interest (e.g. the 5-year recurrence risk), the cumulative/dynamic area under the receiving operator characteristic curve (or AUC_t) can serve as a measure of discrimination.[29] The calculation of AUC_t is similar to the c-index except that patient pairs are only compared if one has a recurrence by 5 years and the other has a recurrence later than 5 years or experiences the competing event (non-recurrence mortality).[30-32] The ordering of two patients having a recurrence after e.g. 2 years and after 3 years will not be in included in this calculation. The AUC_t can be calculated for multiple time-points and shown in a curve.

In the breast cancer data, the c-index calculated for the time range till 5 years of follow up was 0.71 [95% CI 0.66 to 0.76] and the $AUC_{5\ year}$ was 0.72 [95% CI 0.66 to 0.77]. The AUC_t showed a slightly decreasing trend over time with wide confidence intervals (Supplementary Figure 2).

Overall prediction error

Overall model performance entails the overall ability of the model to predict whether a patient does or does not experience the primary event by a particular time point, combining both the calibration and the discrimination of a model. The Brier score summarizes the squared difference between the event indicators and the risk estimates.[33-35] For the competing risks setting, the Brier score is the average squared difference between the primary event indicator at the end of the prediction horizon and the absolute risk estimates by that time-point.[36,18] Weighting techniques or pseudo-observations can account for censoring (Box 1).[36, 37]

The Brier score can range from 0, for a perfect model, to 0.25, for a non-informative model in a dataset with an overall 50% event occurrence. When the overall outcome risk is lower, the maximum score for a non-informative model is lower, which complicates interpretation. Therefore, a scaled version of the Brier score has been proposed: 1-(model Brier score / null model Brier score).[34, 38-40] The null model (without covariates) is a model that estimates the risk equally for all individuals and can in the setting of competing events be estimated by the Aalen-Johansen estimator.[13] The scaled Brier score can be interpreted as an R-squared type of measure, representing the amount of prediction error in a null model that is explained by the prediction model. It has a 'higher is better' interpretation with 100% corresponding to a perfect model, 0% a useless model and <0% a harmful model in the sense that the predictions are further away from the observed data compared to the null model estimating the average risk for each patient.

In the breast cancer validation cohort, the Brier score (lower is better) was 0.09 [95% CI 0.04 to 0.13]. The scaled Brier score (higher is better) showed 5.7% [95% CI 1.6% to 8.2%] explained variation, which we consider fairly low.

Decision curve analysis

Discrimination, calibration and overall prediction error as described above are important when validating a prediction model but do not tell us whether the model would do more good than harm if used in clinical practice.[41,42] To use a risk model for making decisions, we have to choose a risk threshold. Patients with a risk exceeding the threshold are selected for additional clinical interventions. Using the risk model in this way leads to justified interventions (interventions in patients who would develop recurrence) and unnecessary interventions (interventions in patients who would not develop recurrence). The Net Benefit statistic is based on the proportion of justified interventions minus the proportion of unnecessary interventions (Box 2). However, it assigns a weight to the proportion of unnecessary interventions. This weight is related to the chosen threshold: the lower the threshold, the more we value justified interventions and the more we accept unnecessary interventions. The choice of the threshold depends on the (perceived) benefits and harms of the intervention. For example, a highly effective intervention with few side effects suggests using a low threshold. Different clinicians and patients may prefer different thresholds. Therefore, Net Benefit can be calculated for a range of reasonable thresholds, resulting in a decision curve.[41,43] The decision curve of a model is commonly compared to a reference scenario in which all patients receive the intervention ('treat all') and another scenario in which no intervention is given ('treat none').

Box 2: Net Benefit for competing risks data

- Suppose a physician finds it reasonable that, to treat one patient who would otherwise develop a recurrence within 5 years, (e.g. with adjuvant systemic therapy), at most four patients are treated unnecessarily. This means at least 20% of treatments should be justified implying a risk threshold of 20%..
- The benefit of a prediction model is defined as the proportion of patients that are correctly classified as high risk. In presence of competing events, this proportion can be calculated as the cumulative incidence of recurrence among patients with estimated risk at or above 20%, multiplied by the proportion out of all patients with risk at or above 20%.
- The harm from using the model is defined as the proportion of patients who are incorrectly classified as high risk. With competing events, this is calculated as one minus the cumulative incidence among patients with estimated risk exceeding 20% multiplied by the probability of exceeding that threshold (Supplementary material 4).[43]
- The Net Benefit is the benefit minus the harm, in which the harm is assigned a weight. This weight is determined by the risk threshold. Here we find it reasonable that at least 20% (1 in 5) treatments is justified implying that the harm of an unnecessary treatment is considered four times smaller than the benefit of a justified treatment. The weight is therefore 1/4.[41,44,45]

The decision curve in Figure 2 shows the Net Benefit for predicting recurrence within 5 years in the validation data. With a risk threshold of 20% (Box 2), the Net Benefit was 0.014 (Table 2). This net result of 14 out of 1000 patients is made up out of 34 patients whom the prediction model points out correctly as they would develop recurrence if untreated (benefit) versus 81 patients whom the model points out incorrectly and are overtreated (harm). Given the weight of 1/4 implied by the risk threshold (Box 2), this leads to the net result of 34-81/4=14 net more benefiting patients when applying the prediction model to 1000 patients.

A Net Benefit greater than zero and exceeding that of the reference scenarios suggests that the prediction model can add value to clinical decision making. The decision whether or not to implement a model in practice will be further based on practical considerations such as costs and ease with which the information needed in the model can be obtained. In our breast cancer illustration, all four variables are readily available, but in other cases covariate information can be expensive or invasive to obtain. Preferably a formal impact trial is performed to obtain definite evidence on the clinical utility of using a prediction model for clinical decision making.[46]

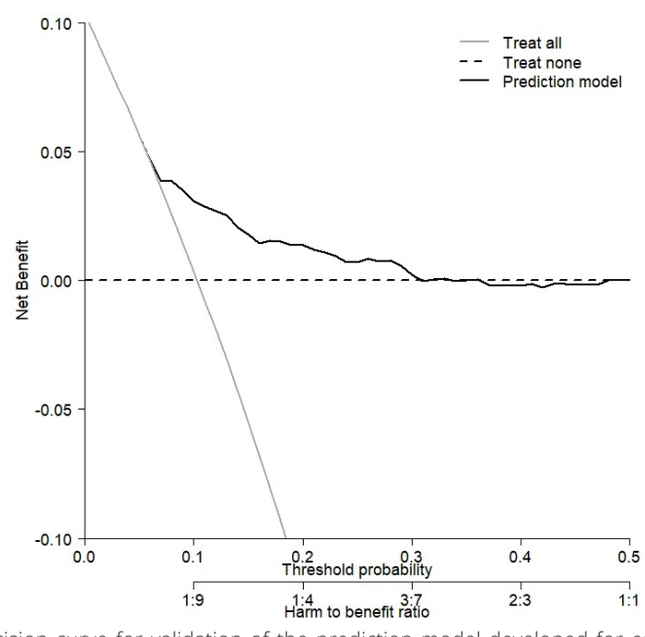


Figure 2: Decision curve for validation of the prediction model developed for estimation of the absolute risk of breast cancer recurrence. The solid black line refers to a scenario where the predictions from the model are compared to the threshold probabilities to decide who receives the intervention. The solid gray line refers to a scenario where all patients receive the intervention. The dashed line refers to a scenario where no patients receive the intervention.

CONCLUDING REMARKS

We provided an overview of performance measures for a comprehensive assessment of the performance of a competing risks prediction model. This typically requires specialist techniques to address censored data such as reweighing the observations or using pseudo-observations. Contemporary, free software facilitates all of the described approaches. The methods can be used for validating any developed time-to-event prediction model, as long as reporting enables calculation of absolute risk estimates for new patients at the time-point(s) of interest.

We recognize that other performance measures are available that have not been described in this overview, which may be important under specific circumstances. For example, methods have been proposed for evaluating estimated absolute risks for several or all competing events at the same time.[47,48] Also, with exception of the c-index and AUC_t curve we limited our descriptions to evaluating absolute risk predictions by a single specific time-point, since this is relevant for most clinical prediction problems. Several of the performance measures that we described can be extended to evaluating predictions by multiple time points or over the entire range of follow-up. Furthermore, we note that large sample sizes are often required for a reliable assessment of performance.[49-51]

The discussed performance measures relate to the full risk distribution (calibration, discrimination, overall performance) and to a decision-analytic perspective (potential impact to obtain better patient outcomes, or clinical utility). These measures are in line with the TRIPOD guidelines, which form a key framework for reporting of prediction models, including the increasingly common competing risks prediction models.[52]

Footnotes

Contributors: All authors provided a substantial contribution to the design and interpretation of the paper and revised drafts. ES initiated the project. NvG wrote the initial draft and is the guarantor for the study. DG analysed the breast cancer data. EB drafted the technical descriptions in Supplementary material 4. DG and EB are main authors of the GitHub page. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: The research of MPP is supported by the Slovenian Research Agency (grant P3-0154, "Methodology for data analysis in medical sciences").

Supplementary material 1: Ignoring competing events
Supplementary material 2: Details on model development
Supplementary material 3: Details on calibration measures
Supplementary material 4: Technical description of the performance measures
Supplementary material 5: Supplementary Table and Figures
Dataset and R code (GitHub page)

REFERENCES

1. Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375. doi:10.1136/bmj.b375
2. Koller MT, Raatz H, Steyerberg EW, et al. Competing risks and the clinical community: irrelevance or ignorance? Statist Med 2012;31:1089–97. doi:10.1002/sim.4384
3. Pfeiffer RM, Gail MH. Absolute risk: methods and applications in clinical management and public health. First issued in paperback. Boca Raton: CRC Press 2020.
4. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 2007;26:2389–430. doi:10.1002/sim.2712
5. Ramspek LR, Teece L, Snell KIE et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. Int J of Epidemiology 2021. https://doi.org/10.1093/ije/dyab256
6. Ramspek CL, Evans M, Wanner C, et al. Kidney Failure Prediction Models: A Comprehensive External Validation Study in Patients with Advanced CKD. JASN 2021;32:1174–86. doi:10.1681/ASN.2020071077
7. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Second edition. Cham, Switzerland: : Springer 2019.
8. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology 2013;13:33. doi:10.1186/1471-2288-13-33
9. Riley RD, Windt D van der, Croft P, et al. Prognosis research in healthcare: concepts, methods, and impact. 2019. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5891544 (accessed 20 Apr 2021).
10. McLernon DJ, Giardiello D, van Calster B, Wynants L, van Geloven N, van Smeden M, Therneau T, Steyerberg EW. Assessing performance in prediction models with survival outcomes: practical guidance. In preparation.
11. Sauerbrei W, Abrahamowicz M, Altman DG, et al. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statist Med 2014;33:5413–32. doi:10.1002/sim.6265
12. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245–7. doi:10.1016/j.jclinepi.2015.04.005
13. Aalen OO, Johansen S. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. Scandinavian Journal of Statistics 1978;5:141–50.https://www.jstor.org/stable/4615704 (accessed 24 Nov 2020).
14. Kattan MW, Giri D, Panageas KS, et al. A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy. Cancer 2004;101:2509–15. doi:10.1002/cncr.20635
15. Wolbers M, Koller MT, Witteman JCM, et al. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. Epidemiology 2009;20:555–61. doi:10.1097/EDE.0b013e3181a39056
16. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Cham Heidelberg New York: : Springer 2015.
17. Andersen PK, Perme MP. Pseudo-observations in survival analysis: Statistical Methods in Medical Research 2010;19(1):71-99. doi: 10.1177/0962280209105020.
18. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing

risks. Statistics in Medicine 2014;33:3191–203. doi:https://doi.org/10.1002/sim.6152
19. Royston P. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. The Stata Journal 2014;14:738–55. doi:10.1177/1536867X1401400403
20. Austin PC, Harrell FE, Klaveren D van. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 2020;39:2714–42. doi:https://doi.org/10.1002/sim.8570
21. Austin PC, Putter H, Giardiello D, et al. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagnostic and Prognostic Research 2022;6:2. doi:10.1186/s41512-021-00114-6
22. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Stat Med 1995;14:1999–2008. doi:10.1002/sim.4780141806
23. Steyerberg EW, Borsboom GJJM, van Houwelingen HC, et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statist Med 2004;23:2567–86. doi:10.1002/sim.1844
24. Harrell FE. Evaluating the Yield of Medical Tests. JAMA 1982;247:2543. doi:10.1001/jama.1982.03320430047030
25. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statist Med 2011;30:1105–17. doi:10.1002/sim.4154
26. Gerds TA, Kattan MW, Schumacher M, et al. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Statist Med 2013;32:2173–84. doi:10.1002/sim.5681
27. Wolbers M, Blanche P, Koller MT, et al. Concordance for prognostic models with competing risks. Biostatistics 2014;15:526–39. doi:10.1093/biostatistics/kxt059
28. Robins JM, Rotnitzky A. Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In: Jewell NP, Dietz K, Farewell VT, eds. AIDS Epidemiology: Methodological Issues. Boston, MA: : Birkhäuser 1992. 297–331. doi:10.1007/978-1-4757-1229-2_14
29. Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of t-year predicted risks. Biostatistics 2019;20:347–57. doi:10.1093/biostatistics/kxy006
30. Saha P, Heagerty PJ. Time-Dependent Predictive Accuracy in the Presence of Competing Risks. Biometrics 2010;66:999–1011. doi:10.1111/j.1541-0420.2009.01375.x
31. Zheng Y, Cai T, Jin Y, et al. Evaluating Prognostic Accuracy of Biomarkers under Competing Risk. Biometrics 2012;68:388–96. doi:10.1111/j.1541-0420.2011.01671.x
32. Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statist Med 2013;32:5381–97. doi:10.1002/sim.5958
33. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Wea Rev 1950;78:1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
34. Graf E, Schmoor C, Sauerbrei W, et al. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 1999;18:2529–45. doi:https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5
35. Gerds TA, Schumacher M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. Biometrical Journal 2006;48:1029–40. doi:https://doi.org/10.1002/bimj.200610301
36. Schoop R, Beyersmann J, Schumacher M, et al. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biom J 2011;53:88–112. doi:10.1002/bimj.201000073
37. Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks models with time-dependent covariates. Statistics in Medicine 2013;32:3089–101. doi:https://doi.org/10.1002/sim.5773
38. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology 2010;21:128–38. doi:10.1097/EDE.0b013e3181c30fb2
39. van Houwelingen H, Putter H. Dynamic Prediction in Clinical Survival Analysis. 0 ed. CRC Press 2011. doi:10.1201/b11311
40. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagn Progn Res 2018;2:7. doi:10.1186/s41512-018-0029-2
41. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74. doi:10.1177/0272989X06295361
42. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;:i6. doi:10.1136/bmj.i6
43. Vickers AJ, Cronin AM, Elkin EB, et al. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak 2008;8:53. doi:10.1186/1472-6947-8-53
44. Kerr KF, Brown MD, Zhu K, et al. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. JCO 2016;34:2534–40. doi:10.1200/JCO.2015.65.5654
45. Pauker SG, Kassirer JP. The Threshold Approach to Clinical Decision Making. N Engl J Med 1980;302:1109–17. doi:10.1056/NEJM198005153022003
46. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013;10:e1001381. doi:10.1371/journal.pmed.1001381
47. Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. Journal of Biomedical Informatics 2015;54:283–93. doi:10.1016/j.jbi.2014.12.016
48. Ding M, Ning J, Li R. Evaluation of competing risks prediction models using polytomous discrimination index. Canadian Journal of Statistics; early view doi:10.1002/cjs.11583
49. Vergouwe Y, Steyerberg EW, Eijkemans MJC, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475–83. doi:10.1016/j.jclinepi.2004.06.017
50. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016;35:214–26. doi:10.1002/sim.6787
51. Pavlou M, Qu C, Omar RZ, et al. Estimation of required sample size for external validation of risk models for binary outcomes. Stat Methods Med Res 2021;30:2187–206. doi:10.1177/09622802211007522
52. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594. doi:10.1136/bmj.g7594

SUPPLEMENTAL MATERIAL

Supplementary material 1 - Ignoring competing risks during model validation

The following results are adapted from Tables 1 and 2 and Figures 3 and 4 published in a study by Ramspek et al., with permission [w1].

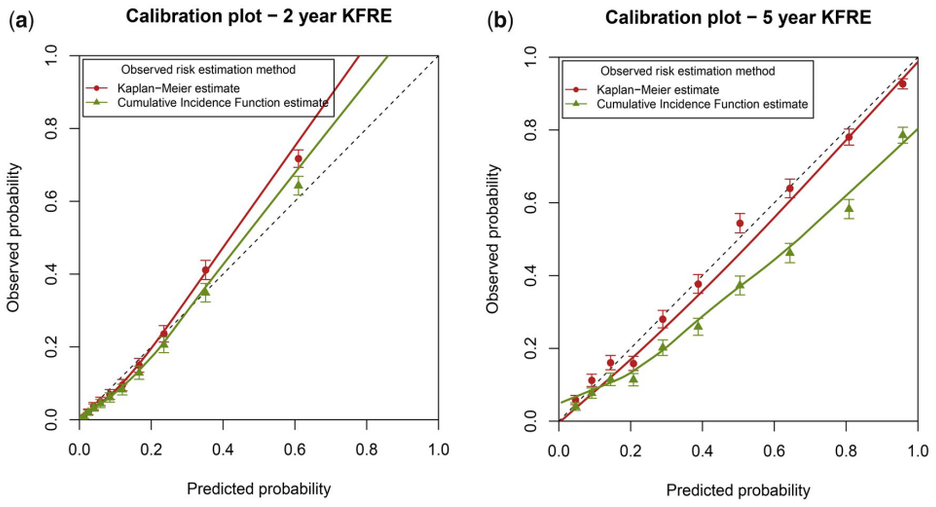


Figure 1: Calibration plots for external validation of the 2- and 5-year Kidney Failure Risk Equation (KFRE). The external validation was performed ignoring competing risks (red points and line) and by using a competing-risks approach (green points and line).

Table 1: Calibration and discrimination results for external validation of the 2- and 5-year KFRE, in the entire validation cohort (n = 13 489). The external validation was performed in two manners, first by ignoring the competing risk of death by censoring these patients and using Kaplan–Meier estimates and second by validating the models whilst taking account of competing risks in the performance measures.

	KFRE 2-year model		KFRE 5-year model	
	Ignoring competing events by censoring	Taking competing events into account	Ignoring competing events by censoring	Taking competing events into account
Average predicted risk	17%	17%	41%	41%
Average observed probability (95% CI)	18% (17%–19%)	16% (15%–17%)	41% (40%–42%)	31% (30%–32%)
O/E ratio (95% CI)	1.06 (1.02–1.10)	0.94 (0.91–0.98)	1.00 (0.98–1.02)	0.76 (0.74–0.78)
C-index (95% CI)	0.840 (0.831–0.849)	0.834 (0.825–0.843)	0.829 (0.821–0.837)	0.814 (0.806–0.822)

KFRE, Kidney Failure Risk Equation; O/E, observed/expected; CI, confidence interval.

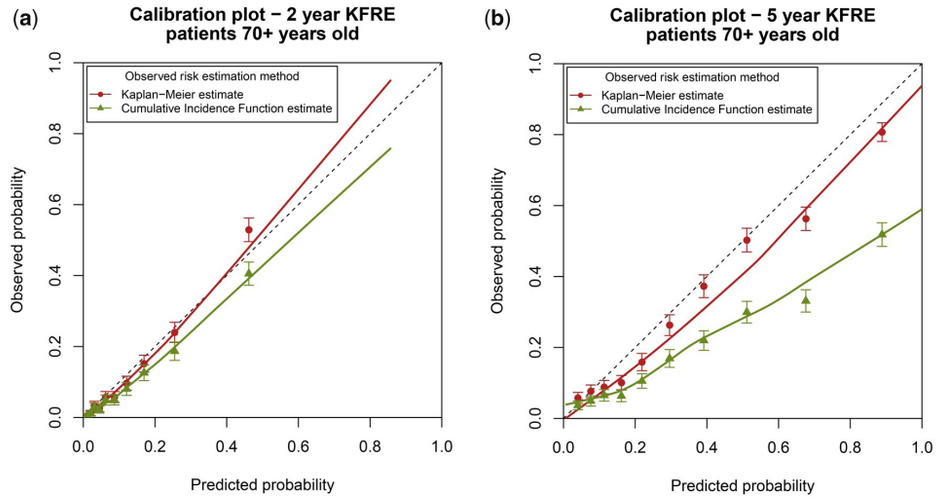


Figure 2: Calibration plots for external validation of the 2- and 5-year Kidney Failure Risk Equation (KFRE) in a subset of older patients. The external validation was performed ignoring competing risks (red points and line) and by using a competing-risks approach (green points and line). The competing-risks approach represents the model performance for the absolute kidney-failure risk in a setting in which patients may die.
In panel (b) the patients with 10% highest risk have an estimated probability of 0.89. when ignoring competing events, the observed outcome probability is 0.81, whereas when accounting for competing events the observed outcome probability is only 0.52 (29 percentage points lower).

Table 2: Calibration and discrimination results for external validation of the 2- and 5-year KFRE, in a subset of patients aged ≥70 years (n = 8654). The external validation was performed in two manners, first by ignoring the competing risk of death by censoring these patients and using Kaplan–Meier estimates and second by validating the models whilst taking account of competing risks in all performance measures.

	KFRE 2-year model		KFRE 5-year model	
	Ignoring competing events by censoring	Taking competing events into account	Ignoring competing events by censoring	Taking competing events into account
Average predicted risk	13%	13%	34%	34%
Average observed probability (95% CI)	11% (11%–12%)	10% (9%–10%)	28% (27%–29%)	19% (18%–20%)
O/E ratio (95% CI)	0.91 (0.86–0.96)	0.78 (0.73–0.83)	0.84 (0.81–0.87)	0.57 (0.54–0.59)
C-index (95% CI)	0.826 (0.810–0.841)	0.813 (0.797–0.828)	0.817 (0.803–0.830)	0.791 (0.778–0.805)

KFRE, Kidney Failure Risk Equation; O/E, observed/expected; CI, confidence interval.

w1 Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. International Journal of Epidemiology 2021:dyab256. doi:10.1093/ije/dyab256

Supplementary material 2 Details on the development of the prediction model

Cause specific versus sub-distribution approach

Analysis methods for predicting absolute risks in competing risks data typically use either the cause-specific hazards for all events (CSH approach) or the sub-distribution hazard of the primary event (SDH approach). In short, in the CSH approach separate regression models are developed for each event, censoring patients who experience the other events. By combining the separate models, the absolute risk of the primary event can be calculated.[w1] In the SDH approach, a single regression model is developed that directly relates to the absolute risk of the primary event.[w2,w3] More details on both approaches can be found in Supplementary material 4.

Although most published competing risks prediction models used the SDH approach (in particular the Fine and Gray model), the CSH approach has two important advantages. Firstly, when calculating absolute risks for multiple competing events, the sum of these risks should remain below one. With the CSH approach this is guaranteed, whereas in the Fine and Gray model it is not.[w4] Secondly, in the CSH approach the hazard ratios are well interpretable as they link to a single event instead of to a combination of events.[w5,w6] This can be useful for understanding a model's behavior and allows including causal thinking into model development which in turn may lead to models that generalize more easily.[w7,w8] Subdistribution (SD) hazard ratios from a Fine and Gray model may be interpreted as directly reflecting the association with absolute risks at the price of a proportionality assumption of such hazard ratios that is difficult to motivate from a biological viewpoint. For instance, a variable may appear protective for the event of interest based on a SD hazard ratio below one, whereas actually it could just as well be a risk factor for the event of interest if the variable is a strong risk factor for a competing event at the same time.

In contrast to the SDH approach, a practical disadvantage of the CSH approach is that calculating absolute risk estimates for new patients cannot be done by hand with a simple formula. It requires access to the cause-specific baseline hazard functions over time up to and including the time point of interest, the cause-specific hazard ratios for each event and the reference levels of the covariates they refer to. As individual patient predictions are typically made through electronic tools (webforms or apps), no issues are foreseen when using such models in clinical practice. For scientific validation of prediction models, the model information is preferably shared in full to facilitate calculating predictions for many new patients in one go. We provide R code for sharing and using model information when using the CSH approach without having to share raw data at our GitHub page. For the SDH approach, calculating the absolute risks for new patients requires the estimated baseline absolute cumulative risk at the prediction horizon, the sub-distribution hazard ratios for the primary event and the reference levels of the covariates that they refer to.

The prediction model we use for illustration of performance measures in the manuscript was developed using the CSH approach. The discussed validation methods are equally applicable to other competing risks models such as the SDH approach, (flexible) parametric models and random survival forests, as long as the models provide sufficient information to calculate the estimated absolute risks for new patients.

Development data

We developed the prediction model on the FOCUS cohort.[w9] In this retrospective cohort, all consecutive patients aged 65 years or older with breast cancer diagnosed in the South-West region of the Netherlands in the years 1997-2004 were included. The registry contains information on patient-characteristics including tumor characteristics, treatment and disease recurrence. Follow-up data on patient survival (maximal 5 years) was obtained by linkage with the municipal population registries. We applied the following inclusion criteria (same inclusion criteria that were used in the validation cohort): patients with primary breast cancer who received primary breast surgery, and received no previous neoadjuvant treatment. We used a random subset of 1000 patients to allow Open Access data sharing. Out of these 1000 patients in the development set, 135 developed breast cancer recurrence and 204 had a non-recurrence death within the five years follow up (cumulative incidence curve in Supplementary Figure 1). Except for the higher age inclusion criterion in the validation cohort, patients were rather similar on the listed characteristics in the development and validation cohorts (Supplementary Table 1).

Model development

Using the CSH approach, we combined the two Cox proportional hazards models for recurrence and death. In both models, we used age, tumor size, nodal status, and hormone receptor status as predictors. We assessed the proportionality assumptions of the models visually and with tests based on Schoenfeld residuals, and did not observe strong deviations. We assessed the linearity of the effects of age and tumor size by comparing model fit (Akaike's Information Criterion) using linear covariate effects and using restricted cubic splines. Linear effects showed adequate fit. Larger tumor size, positive nodal status and negative hormone receptor status were strong predictors of breast cancer recurrence (Supplementary Table 2). Age was strongly related to non-recurrence mortality.

Fine and Gray model

For completeness we repeated our illustration with a model developed using the SDH approach. Code for development and validation of such a model is available from out GitHub page. In the SDH approach, we used a Fine and Gray sub-distribution hazards model following the same steps as in the CSH approach. Validation results were highly similar to those of the CSH approach presented in the main manuscript.

w1 Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 2007;26:2389–430. doi:10.1002/sim.2712

w2 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association 1999;94:496–509. doi:10.2307/2670170

w3 Gerds TA, Scheike TH, Andersen PK. Absolute risk regression for competing risks: interpretation, link functions, and prediction. Statistics in Medicine 2012;31:3921–30. doi:10.1002/sim.5459

w4 Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. Statistics in Medicine 2021;40:4200–12. doi:10.1002/sim.9023

w5 Lau B, Cole SR, Gange SJ. Competing Risk Regression Models for Epidemiologic Data. American Journal of Epidemiology 2009;170:244–56. doi:10.1093/aje/kwp107

w6 Koller MT, Raatz H, Steyerberg EW, et al. Competing risks and the clinical community: irrelevance or ignorance? Statist Med 2012;31:1089–97. doi:10.1002/sim.4384

w7 Piccininni M, Konigorski S, Rohmann JL, et al. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. BMC Med Res Methodol 2020;20:179. doi:10.1186/s12874-020-01058-z

w8 van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol 2020;35:619–30. doi:10.1007/s10654-020-00636-1

w9 de Glas NA, Kiderlen M, Bastiaannet E, et al. Postoperative complications and survival of elderly breast cancer patients: a FOCUS study analysis. Breast Cancer Res Treat 2013;138:561–9. doi:10.1007/s10549-013-2462-9

Supplementary material 3 Details on calibration measures

Alternative numerical summaries of overall calibration (calibration-in-the-large)

In the main paper we present the O/E ratio to summarize overall calibration into a single number. An alternative way to summarize overall calibration is by calculating the average distance between the calibration curve and the diagonal (i.e., the line that would indicate perfect calibration). When the distance is averaged on the squared scale, this leads to what has been referred to as the 'mean squared bias'.[w1,w2] When reported, we recommend using the root mean squared bias to facilitate interpretation. To calculate the distance between the calibration curve and diagonal, we need the (smoothed) estimate of the observed outcome proportion for each patient's estimated risk. As for the calibration curve, these smoothed outcome proportions can be estimated using pseudo-observations[w1] or by using a flexible regression model[w3,w4] and will depend on the chosen degree of smoothing (Box 1). The difference with the definition of the Brier score discussed in the main of the paper is that we here compare the predictions to observed outcome proportions, and not to individual (zero or one) primary event indicators as is the case with the Brier score.

Recently, averaging the distance on the absolute scale was proposed, leading to a measure called the integrated calibration index (ICI).[w3,w4] Both the root mean squared bias and the ICI indicate how far off target the risk estimates are on average. We prefer averaging on the squared scale as previous literature has pointed out that absolute distance measures in the survival setting may lack a desired statistical property called 'propriety', meaning that a perfect model that provides the true underlying risks does not necessarily score best.[w5] In line with earlier work, we propose also reporting the median (E50) and 90th percentile (E90) of the absolute differences along with ICI and/or root mean squared bias.[w6]

Results from the breast cancer validation cohort are presented in the table below. The root mean squared bias and ICI show that on average the model was 3 percentage points off target, with 90% of observations staying within 5 percentage points error.

Table: Estimated values of the additional measures for overall calibration in the external breast cancer data

Root mean squared bias	0.035
ICI	0.031
E50	0.030
E90	0.052
Emax	0.159

Calibration intercept and calibration slope for competing risks data

A pseudo-observation is used as a proxy measure of the primary event indicator at the time-point of interest for each patient (did the patient experience the primary event before or at the prediction horizon or not). The pseudo-observations are calculated as the weighted difference between the cumulative incidence estimate at the prediction horizon based on all patients and the same quantity estimated leaving that patient out. These are so-called 'jackknife' pseudo-observations. Note that these individual pseudo-observations can have unintuitive values beyond the 0-1 range and may even be negative. The important property of pseudo-observations that is employed when they are used for assessment of calibration is that on average they give an unbiased estimate of the observed cumulative incidence.[w7, w2]. Similar to the setting of ordinal time-to-event outcomes, to calculate calibration intercept and slope, the pseudo-observations can be regressed using a generalized linear model with (a complementary log-log transformation of) the risk estimates as an offset, meaning that the regression coefficient of the risk estimates is constrained to one.[w8] The estimated intercept from this model is the calibration intercept and indicates how much the risk estimates are over- or underestimating on average. A negative calibration intercept indicates that the risk estimates are on average too high and a positive intercept indicates that the risk estimates are on average too low. The calibration slope can be estimated by adding (on top of the offset described above) the same (complementary log-log transformed) risk estimates as a covariate to the generalized linear model. The estimated regression coefficient for this covariate indicates how much the calibration slope deviates from one. A calibration slope between 0 and 1 indicates too extreme predictions of the model, i.e. for patients with low risks the estimated risks are too low and for patients with high risk the estimated risks are too high. A calibration slope >1 indicates predictions do not show enough variation. A calibration slope <0 would imply that predictions are in the wrong direction.

The calculations can be extended from risk up to one particular time-point to a calibration intercept and slope that are based on a range of time points spanning the follow-up period.[w8]

Alternatively, if focus is not on a single time point but on the full range of observed follow up, a calibration intercept and slope could be estimated by a procedure using Poisson regression.[w9, w10]

w1 Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks models with time-dependent covariates. Statistics in Medicine 2013;32:3089–101. doi:https://doi.org/10.1002/sim.5773

w2 Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. Statistics in Medicine 2014;33:3191–203. doi:https://doi.org/10.1002/sim.6152

w3 Austin PC, Putter H, Giardiello D, et al. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagnostic and Prognostic Research 2022;6:2. doi:10.1186/s41512-021-00114-6

w4 Austin PC, Harrell FE, Klaveren D van. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 2020;39:2714–42. doi:https://doi.org/10.1002/sim.8570

w5 van Houwelingen H, Putter H. Dynamic Prediction in Clinical Survival Analysis. 0 ed. CRC Press 2011. doi:10.1201/b11311

w6 Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Cham Heidelberg New York: : Springer 2015.

w7 Andersen PK. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 2003;90:15–27. doi:10.1093/biomet/90.1.15

w8 Royston P. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. The Stata Journal 2014;14:738–55. doi:10.1177/1536867X1401400403

w9 Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res 2016;25:1692–706. doi:10.1177/0962280213497434

w10 Brentnall AR, Cuzick J. Risk Models for Breast Cancer and Their Validation. Stat Sci 2020;35:14–30. doi:10.1214/19-STS729

Supplementary material 4: Technical description of the performance measures

1. General notation

We use the tutorial paper by Putter et al. [1] as main reference for the Sections 1 through 3. We assume that individuals can experience one of *K* distinct events. We denote the failure time as *T*, and the competing event indicator as $D \in \{1, ..., K\}$. In practice, individuals are subject to some right-censoring time *C*, which is assumed to be independent of *T* and *D*, possibly given covariates. We thus only observe realizations of $\tilde{T} = \min(C, T)$ and $\tilde{D} = I(T \leq C)D$, where $\tilde{D} = 0$ indicates a right-censored observation and $I(\cdot)$ is the indicator function.

2. Key quantities

The *cause-specific hazard* of failing from a cause *k* in presence of competing events is defined as:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t}.$$

The overall survival probability is defined by the *K* cause-specific hazard functions as

$$S(t) = \exp\left(-\sum_{k=1}^{K} \int_0^t h_k(u)du\right) = \exp\left(-\sum_{k=1}^{K} H_k(t)\right),$$

Where $H_k(t) = \int_0^t h_k(u)du$ is the cause-specific cumulative hazard for cause *k*.
The *cumulative incidence function* for an event *k*, also referred to as the absolute risk of event *k*, is the probability of that event occurring by a particular time-point *t* without any other competing event occurring earlier, $P(T \leq t, D = k)$. It is defined as

$$F_k(t) = \int_0^t h_k(u)S(u-)du,$$

with $S(u-)$ being the total survival probability just up to time *u*.

3. Aalen-Johansen

Suppose we observe *n* independent samples $(\tilde{t}_i, \tilde{d}_i)$ of $(\tilde{T}, \tilde{D})$, for $i = 1 \cdots n$. We order the *J* distinct event times where any of the *K* competing events occur as $0 < t_1 < ... < t_J$. Let $D_k(t_j)$ denote the number of individuals failing from cause *k* at t_j, and let $D(t_j) = \sum_{k=1}^{K} D_k(t_j)$ denote the total number of failures from any cause at t_j. The number of individuals at risk of any event at t_j is given by $R(t_j)$.

The cumulative incidence of cause *k* by some time horizon *s* can be estimated non-parametrically using the Aalen-Johansen estimator [2], defined as

$$\widehat{F}_k(s) = \sum_{j:t_j \leq s} \widehat{h}_k(t_j)\widehat{S}(t_{j-1}),$$

Where

$$\widehat{h}_k(t_j) = \frac{D_k(t_j)}{R(t_j)}, \qquad \widehat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \sum_{k=1}^{K} \widehat{h}_k(t_j)\right).$$

This Aalen-Johansen estimator is sometimes referred to directly as 'the cumulative incidence function' (e.g. Ramspek 2021+). Here we the denote the cumulative incidence function as the population quantity we are targeting, and the Aalen-Johansen estimator as the means to estimate it from data.

4. Regression models

We assume for the remainder of this document that primary interest lies in estimating the cumulative incidence for event *D* = 1 by some prediction horizon *s*, conditional on covariates. Let **Z** denote a vector of *p* covariates, which are observed for every i^{th} individual as $\mathbf{z}_i$.

The two most commonly used methods for predicting an event conditional on covariates in the presence of competing risks are the Fine and Gray approach [3], and the cause-specific Cox proportional hazards approach. Both are able to produce a subject-specific absolute risk of experiencing event *D* = 1 by *s*, which we denote as $\pi_1(s \mid \mathbf{z}_i)$. This is effectively an estimate of $F_1(s \mid \mathbf{z}_i) = P(T \leq s, D = 1 \mid \mathbf{z}_i)$.

4.1 Cause-specific Cox proportional hazards

The cause-specific approach first entails specifying a Cox proportional hazards model for each of the *K* competing events as

$$h_k(t \mid \mathbf{Z}) = h_{k0}(t) \exp(\boldsymbol{\beta}_k^\intercal \mathbf{Z}),$$

where $h_{k0}(t)$ is the cause-specific baseline hazard, and β_k represents the effects of covariates **Z** on the cause-specific hazard. Each model can be estimated by treating all events by causes other than *D* = *k* as censored. Note that the models need not necessarily share the same covariates.

In order to obtain $\pi_1(s \mid \mathbf{z}_i)$ using the cause-specific approach, the individual-specific hazards must first be calculated as

$$\widehat{h}_k(t \mid \mathbf{z}_i) = \widehat{h}_{k0}(t) \exp(\hat{\boldsymbol{\beta}}_k^\intercal \mathbf{z}_i),$$

where $\widehat{h}_{k0}(t)$ is calculated based on the increments in the Breslow estimate of the cause-specific cumulative baseline hazard. These hazards for all *J* distinct timepoints can thereafter be plugged into the formula for $\widehat{F}_k(s)$ outlined in Section 3, producing

$\pi_1(s \mid \mathbf{z}_i)$ for *D* = 1. We refer the reader for example to Section 5.2.1 of the text by Beyersmann et al. [4] for a more detailed treatment of the procedure.

4.2 Fine and Gray approach

The Fine and Gray approach uses a model for the so-called *subdistribution hazard*, defined for cause *D* = *k* as

$$\begin{aligned}\lambda_k(t \mid \mathbf{Z}) &= \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t, D = k \mid T \geq t \cup (T \leq t \cap D \neq k), \mathbf{Z}\}}{\Delta t}, \\ &= \frac{-d \log\{1 - F_k(t \mid \mathbf{Z})\}}{dt},\end{aligned}$$

where patients failing from competing causes $D \neq k$ remain in the risk-set up to the end of follow up.

A proportional hazards model can be specified for this subdistribution hazard as

$$\lambda_k(t \mid \mathbf{Z}) = \lambda_{k0}(t) \exp(\boldsymbol{\gamma}_k^{\intercal} \mathbf{Z}),$$

with $\lambda_{k0}(t)$ being the subdistribution baseline hazard function and γ_k representing the effects of covariates **Z** on the subdistribution hazard. The cumulative incidence function for *D* = *k* can then be written as

$$F_k(s \mid \mathbf{Z}) = 1 - \exp\left[- \exp(\boldsymbol{\gamma}_k^{\intercal} \mathbf{Z}) \int_0^s \lambda_{k0}(u) du \right],$$

or equivalently,

$$1 - F_k(s \mid \mathbf{Z}) = \{1 - F_{k0}(s)\}^{\exp(\boldsymbol{\gamma}_k^{\intercal} \mathbf{Z})},$$

where $F_{k0}(s)$ denotes the baseline cumulative incidence. Thus, for event *D* = 1 this model can be used directly to obtain a prediction $\pi_1(s \mid \mathbf{z}_i)$ without having to model the other competing causes.

5 Dealing with censoring when assessing performance

Let T_i and D_i respectively denote the true event time and competing event indicator for an individual *i*. We can define $Y_i(s) = I(T_i \leq s, D_i = 1)$ as the binary event which indicates whether event *D* = 1 occurred prior to the prediction horizon *s*, or not. If an individual *i* is censored prior to *s*, we cannot know whether they would have gone on to experience the event of interest or not. Hence, $Y_i(s)$ is not fully observed in the presence of right-censoring.

5.1 Pseudo-observations

One of the ways to deal with the issue of censoring is to use pseudo-observations $\tilde{Y}_i(s)$ [5], which attempts to recreate $Y_i(s)$. These are defined as

$$\tilde{Y}_i(s) = n\widehat{F}_1(s) - (n-1)\widehat{F}_1^{-i}(s)$$

where $\widehat{F}_1(s)$ is the Aalen-Johansen estimate of $\mathbb{E}\{Y_i(s)\}$ based on all patients, and $\widehat{F}_1^{-i}(s)$ is based on the sample excluding the i^{th} individual. In case of covariate-dependent censoring, a weighted version of the Aalen-Johansen estimator should instead be used [6]. Using $\tilde{Y}_i(s)$ instead of $Y_i(s)$ in the calculation of for instance performance measures eases up calculations as all individuals have a value for $\tilde{Y}_i(s)$.

5.2 IPCW

Another way to deal with the issue of censoring is to use inverse probability of censoring weights (IPCW). Individuals with an observed event status at *s* are known as a 'complete-case', meaning they have either experienced one of *K* events prior to *s*, or are still at risk at *s*. Conditional on covariates $\mathbf{z}_i$ and experiencing an event at $\tilde{t}_i \leq s$, the probability of still being under follow-up just prior to $\tilde{t}_i$ is denoted by $G(\tilde{t}_i - \mid \mathbf{z}_i)$. For those still at risk, $\tilde{t}_i > s$, the probability of being observed to have no event up to time s is written as $G(s \mid \mathbf{z}_i)$. Both can be estimated using the Cox proportional hazards models, or by Kaplan-Meier estimators in absence of any $\mathbf{z}_i$ predictive of censoring.

Individuals who are known to have experienced a particular event before time *s* or to still be at risk prior to *s* are then weighted inversely to their probability of having that particular outcome, $1/G(\tilde{t}_i - \mid \mathbf{z}_i)$ or $1/G(s \mid \mathbf{z}_i)$.

6 Performance measures

6.1 Calibration

As per Blanche et al. [7], strong model calibration is defined by

$$\pi_1(s \mid \mathbf{Z}) = P\{Y(s) = 1 \mid \mathbf{Z}\} \qquad \text{for all } \mathbf{Z},$$

meaning that the estimated risk is equal to the observed outcome proportion for all values (and thus combinations) of **Z**. Unless **Z** is low-dimensional and made up entirely of categorical variables, this is typically impossible to assess. We can instead calibration by means of various graphical and numerical summaries.

6.1.1 Calibration plot

The simplest calibration plot bins individuals into approximately equally sized groups based on their risk estimates, and plots the relationship between the average estimated risk and the observed outcome proportion of the event in *each* group. The latter can be either estimated using the Aalen-Johansen estimator, or by averaging across the

pseudo-observations within a group. Formally, the calibration plot assesses

$$P\{Y(s) = 1 \mid \pi_1(s \mid \mathbf{Z}) = r\} = r \qquad \text{for all } \mathbf{Z}, \text{ for all } r \in [0, 1],$$

which essentially states that among individuals with an estimated risk of *r*, the observed outcome proportion should also be *r*. Methods that attempt to create a *smooth* calibration curve, be it through local smoothing of pseudo-observations [8] or spline-based regression of risk estimates [9], try to create continuity. In other words, they try to make the groups defined by *r* as small as possible.

Briefly, the *subdistribution model* approach (Austin et al. 2020+) to creating a smooth calibration curve fits a Fine and Gray model for the primary event as a flexible function of the estimated risks, which have been transformed as $\log(-\log(1 - \pi_1(s \mid \mathbf{z}_i)))$. Restricted cubic splines are used as the flexible function, where the number of internal *knots* define the degree of smoothing. The predictions from this flexible subdistribution model by *s* serve as the observed outcome proportions, and can be plotted against $\pi_1(s \mid \mathbf{z}_i)$ to create the calibration curve.

The approach taken in [8] to create smooth calibration curves first relies on computing the pseudo-observation $\tilde{Y}_i(s)$ for each individual. Then, for some probability *p*, the pseudo-observations of individuals with an estimated risk within some intervals of *p* are averaged to obtain an observed outcome proportion. This pre-specified interval around *p*, or *bandwidth*, defines the degree of smoothing.

6.1.2 Numerical summaries of calibration

Calibration 'in the large' is defined by

$$\mathbb{E}\{\pi_1(s \mid \mathbf{Z})\} = P\{Y(s) = 1\},$$

stating that the average estimated risk equals the overall observed outcome proportion. A popular way of summarizing this is the ratio of cumulative observed over expected events, or *O/E*. Due to censoring in the current setting, we divide risks instead of absolute event numbers. The observed outcome proportion ('observed') is given by the Aalen-Johansen estimator, while the expected risk is simply the average across all estimated risks. For an alternative calculation of the *O/E* ratio, see [10].

A second type of numerical summary is the integrated calibration index (ICI), which is a weighted mean of the absolute differences between estimated risks and observed outcome proportions [9]. Specifically, let *x* represent the vector of estimated risks $\pi_1(s \mid \mathbf{Z})$ by time *s*, and x_c the value of the calibration curve (i.e. the observed outcome proportions, obtained by smoothing) at *x*. If we define $f(x) = |x - x_c|$, and define the density function of x as $\phi(x)$, then

$$ICI(s) = \int_0^1 f(x)\phi(x)dx,$$

which is estimated as simply the empirical mean of $f(x)$. The median (E50) and or other percentiles of the *f*(*x*) are also possible numerical summaries. Similarly, the squared bias may be of interest, which is estimated as the empirical mean of $f(x) = (x - x_c)^2$.

Note that these numerical summaries depend on the degree and type of smoothing applied to obtain x_c. With higher flexibility, i.e. smaller bandwidth for smoothing the pseudo-observations or higher number of knots in the subdistribution approach, the calibration curve may be overfitted in areas with few observations where the estimated risks are usually very small or large. The advice for the subdistribution approach is to use between 3 and 5 internal knots (Austin et al. 2020+), while for the pseudo-observation approach ample advice is provided in the text by Gerds et al. [8]. Finally, note that the smoothing method chosen to obtain the calibration plot should preferably be the same as the one used when computing the numerical summaries.

A third way to numerically summarize calibration is through the calibration intercept and calibration slope, which additionally allow for miscalibration testing. We briefly explain the extension of the methods described in [11] to the competing risks setting. The idea is to model the pseudo-observations $\tilde{Y}_i(s)$ as a function of the complementary log-log transformed estimated risks $\text{cloglog}\{\pi_1(s \mid \mathbf{z}_i)\} = \log(-\log(1 - \pi_1(s \mid \mathbf{z}_i)))$ in a generalized linear regression model (GLM). By writing $\mathbb{E}\{\tilde{Y}(s)\} = \mu$, we can formulate the following two regression models,

$$\text{cloglog}(\mu) = \beta_0 + \text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}, \qquad (1)$$
$$\text{cloglog}(\mu) = \beta_0' + \beta_1' \,\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}. \qquad (2)$$

Both GLMs use a complementary log-log link function for the mean, and assume constant variance. Additionally, both models are fitted by means of generalized estimating equations (GEE) [12]. Model (1) allows estimation of the calibration intercept β_0, which should ideally be equal to zero. In this model, the transformed risk estimates $\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}$ are used as an *offset*, meaning that its coefficient is constrained to unity. A calibration intercept (significantly) below or above zero respectively implies on average over and underestimation of the observed outcome proportions.

Model (2) allows estimation of the calibration slope β_1', which should ideally be equal to one. A calibration slope between 0 and 1 indicates too extreme predictions (both on the low and on the high side), while a calibration slope greater than 1 indicates predictions that do not show enough variation. A negative calibration slope implies predictions are in the wrong direction. Furthermore, adding the transformed risk estimates as an offset

in model (2) allows to test $\beta_1' = 1$ directly.

Regarding testing, it is preferable to first perform a joint test $(\beta_0', \beta_1') = (0, 1)$with two degrees of freedom to assess overall evidence for miscalibration [13]. If the null-hypothesis is rejected in the joint test, the individual tests for β_0 and β_1' can then be performed.

6.2 Discrimination

We introduce a pair of individuals *i* and *j* with covariates $\mathbf{z}_i$ and $\mathbf{z}_j$ respectively. At horizon *s*, we have model-based predictions $\pi_1(s \mid \mathbf{z}_i)$ and $\pi_1(s \mid \mathbf{z}_j)$. The ordering of these estimated risks at *s* is thus denoted by

$$Q_{ij}(s) = I\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j)\}.$$

6.2.1 C-index

As described in [14], the 'truncated' concordance index (C-index) is defined by

$$\mathcal{C}_1(s) = P\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j) \mid D_i = 1, T_i \leq s, (T_i < T_j \cup D_j \notin \{0, 1\})\}.$$

It measures how well the model ranks the event times occurring prior to *s* [15]. Notice that for a pair of individuals, if the individual with the earlier event time is right-censored, the ordering $T_i < T_j$ is indeterminable. A simple solution for estimating the C-index is setting the follow up time of the patients with competing event to the maximum follow up time in the study design [16]. This method can however only be used in settings without censoring or with purely administrative censoring, as recently illustrated for prediction of kidney failure (Ramspek et al., 2020+). Hence, to estimate the C-index in the presence of other types of right-censoring, we can construct weights as part of an IPCW procedure, yielding

$$w_{ij,1} = \frac{I(\tilde{t}_i < \tilde{t}_j)}{\widehat{G}(\tilde{t}_i - \mid \mathbf{z}_i)\widehat{G}(\tilde{t}_i \mid \mathbf{z}_j)}, \qquad w_{ij,2} = \frac{I(\tilde{t}_i \geq \tilde{t}_j, \tilde{d}_j \notin \{0, 1\})}{\widehat{G}(\tilde{t}_i - \mid \mathbf{z}_i)\widehat{G}(\tilde{t}_j - \mid \mathbf{z}_j)}.$$

We can then estimate the c-index as

$$\widehat{\mathcal{C}}_1(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n (w_{ij,1} + w_{ij,2}) Q_{ij}(s) I(\tilde{t}_i \leq s, \tilde{d}_i = 1)}{\sum_{i=1}^n \sum_{j=1}^n (w_{ij,1} + w_{ij,2}) I(\tilde{t}_i \leq s, \tilde{d}_i = 1)}.$$

We note that the c-index is not appropriate for validating prediction models with time-varying covariate effects [17].

6.2.2 Time-dependent area under the ROC curve

We define cases as individuals with $\tilde{t}_i \leq s$ and $\tilde{d}_i = 1$, i.e. as experiencing the primary event by *s*. Controls however have been defined in two ways:

1. free of any event by s, i.e. $\tilde{t}_i > s$,
2. free of any event by s, i.e. $\tilde{t}_i > s$, or experiencing a competing event, $(\tilde{t}_i \leq s,\ \tilde{d}_i \notin \{0, 1\})$.

We continue with the second definition here. We define a time-dependent area under the receiving operating characteristic curve (AUC_t), described in [18] and the supplementary material of [14].

It is defined as

$$AUC_1(s) = P\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j) \mid D_i = 1, T_i \leq s, (T_j > s \cup D_j \notin \{0, 1\})\}.$$

It evaluates the concordance of risk estimates between individuals experiencing the primary event by *s*, and individuals either event-free or that have experienced a competing event. Similarly to the C-index, a pair becomes unevaluable (directly) if one of the individuals has a right-censored event time prior to *s*. Specifically, we cannot determine whether this individual would experience the primary event between the right-censoring time and *s*, or remain a control. Thus, we must first construct weights

$$w_i = \frac{I(\tilde{t}_i \leq s, \tilde{d}_i = 1)}{\hat{G}(\tilde{t}_i)}, \qquad w_{j,1} = \frac{I(\tilde{t}_j \leq s, \tilde{d}_j \notin \{0, 1\})}{\hat{G}(\tilde{t}_j)}, \qquad w_{j,2} = \frac{I(\tilde{t}_j > s)}{\hat{G}(s)},$$

and then can estimate $AUC_1(s)$ as

$$\widehat{AUC}_1(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_i (w_{j,1} + w_{j,2}) Q_{ij}(s)}{\sum_{i=1}^n w_i \sum_{j=1}^n (w_{j,1} + w_{j,2})}.$$

We refer to [18] for details on covariate dependent censoring.

Alternative versions of the AUC_t have been proposed which use different definitions of cases and controls according to having their events before, at or after the time-point of interest [19]. The cumulative case/dynamic control definition we describe here can be considered most suited for evaluation of predictions from baseline over a specific prediction horizon [20] whereas the incident case/dynamic control definition with cases defined as having the primary event exactly at (i.e. not before) a fixed time-point, can be useful in evaluating dynamic prediction models [20, 21, 22].

6.3 Overall prediction error

6.3.1 Brier score

The Brier score in the context of competing events is the expected quadratic distance between the event indicator $Y(s)$ (for the primary event $D = 1$) and the estimated risks $\pi_1(s \mid \mathbf{Z})$ based on the prediction model,

$$B_1(s) = \mathbb{E}\big[I(T \leq s, D = 1) - \pi_1(s \mid \mathbf{Z})\big]^2,$$

with $I(T \leq s, D = 1)$ being the true event status at *s*. In the presence of censoring, the Brier score can be estimated using either IPCW, or pseudo-observations. The latter estimator has only been suggested in the context of dynamic prediction [23], and so it is

not included in this overview.

As per Schoop et al. [24], an IPCW estimator for the Brier score is

$$\widehat{B}_1(s) = \frac{1}{n}\sum_{i=1}^{n}\left[I(\tilde{t}_i \le s, \tilde{d}_i = 1) - \pi_1(s \mid \mathbf{z}_i)\right]^2 w_{1i},$$

Where

$$w_{1i} = \frac{I(\tilde{t}_i \le s, \tilde{d}_i \ne 0)}{\widehat{G}(\tilde{t}_i - \mid \mathbf{z}_i)} + \frac{I(\tilde{t}_i > s)}{\widehat{G}(s \mid \mathbf{z}_i)}.$$

6.3.2 Scaled Brier Score

As per Kattan and Gerds [25], the scaled Brier score (also know as index of prediction accuracy, IPA) for estimating the cumulative incidence of event *D* = 1 is

$$\mathrm{IPA}(s) = 1 - \frac{B_1^{\mathrm{mod}}(s)}{B_1^{\mathrm{null}}(s)},$$

where $B_1^{\mathrm{mod}}(s)$ is the model Brier score, and $B_1^{\mathrm{null}}(s)$ is the Brier score for the null model (with no covariates). The latter can be calculated by plugging-in the Aalen-Johansen estimator in place of $\pi_1(s \mid \mathbf{z}_i)$.

6.4 Decision curves

In a competing risks setting, the net benefit at *s* based on a prediction model for the primary event, given a chosen probability threshold p_s, is given by

$$\mathrm{NB}_1(s) = \frac{\mathrm{TP}_1(s)}{n} - \frac{\mathrm{FP}_1(s)}{n}\left(\frac{p_s}{1-p_s}\right), \qquad (3)$$

where $\mathrm{TP}_1(s)$ is the true positive count and $\mathrm{FP}_1(s)$ the false positive count.
In order to estimate the net benefit, we first define $x_i = 1$ if $x_i = 1 \text{ if } \pi_1(s \mid \mathbf{z}_i) \ge p_s$. In other words, x_i defines whether an individual is classified as their estimated risk exceeding the chosen probability threshold p_s. $P(X = 1)$ is then the proportion classified as $X = 1$ based on this threshold.

Recall $\widehat{F}_1(s)$ as the Aalen-Johansen estimate of the cumulative incidence of event *D* = 1 by horizon *s*. The quantity is $\widehat{F}_1(s \mid X = 1)$ then the estimated cumulative incidence *among* those classified as exceeding the risk threshold. As described in [26], the number of true positives is estimated as

$$\widehat{\mathrm{TP}}_1(s) = \widehat{F}_1(s \mid X = 1) \times P(X = 1) \times n,$$

and similarly, the number of false positives as

$$\widehat{\mathrm{FP}}_1(s) = \left[1 - \widehat{F}_1(s \mid X = 1)\right] \times P(X = 1) \times n.$$

The estimated net-benefit $\widehat{\mathrm{NB}}_1(s)$ can then be obtained by plugging-in $\widehat{\mathrm{TP}}_1(s)$ and into $\widehat{\mathrm{FP}}_1(s)$ (3). Furthermore, a *decision curve* can be obtained by plotting $\widehat{\mathrm{NB}}_1(s)$ for various values of p_s. This is often plotted alongside a 'treat-all' curve, which plots the net-benefit across thresholds in a situation where all individuals are classified as exceeding the risk threshold regardless of the prediction model. A 'treat none' reference line is useful as well, with net-benefit of zero for any threshold (no true positive and no false positive decisions are made).

7. Closing remarks

Formulas concerning standard errors of performance measures are beyond the scope of this document. Analytical formulas are available for many measures, and bootstrapping can be used for most.

7

REFERENCES

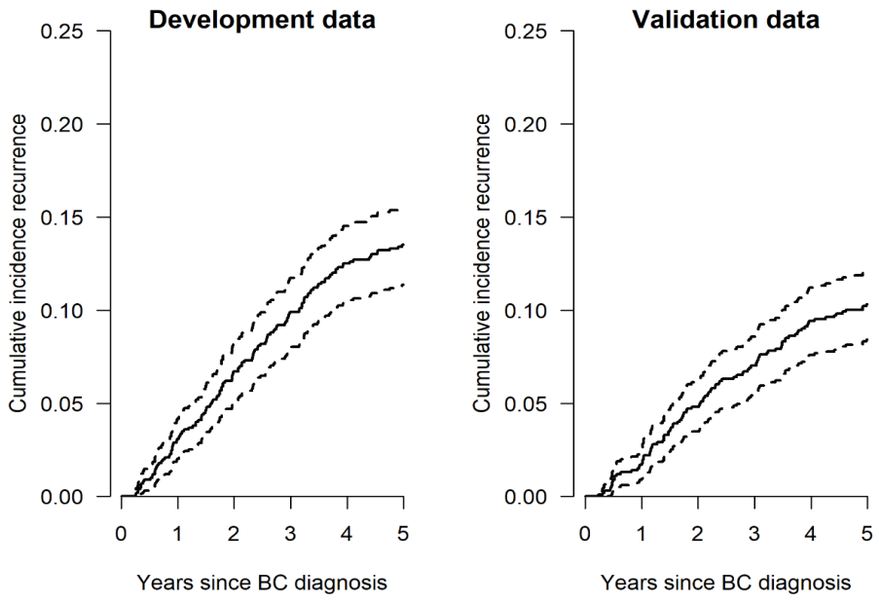
1. H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. ISSN 1097-0258.
2. Odd O. Aalen and Søren Johansen. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3): 141–150, 1978. ISSN 0303-6898.
3. Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. ISSN 0162-1459.
4. Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing Risks and Multistate Models with R*. Use R! Springer-Verlag, New York, 2012. ISBN 978-1-4614-2034-7.
5. Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis:. *Statistical Methods in Medical Research*, August 2009.
6. Nadine Binder, Thomas A. Gerds, and Per Kragh Andersen. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, 20(2):303–315, April 2014. ISSN 1572-9249.
7. Paul Blanche, Thomas A. Gerds, and Claus T. Ekstrøm. The Wally plot approach to assess the calibration of clinical prediction models. *Lifetime Data Analysis*, 25(1):150–167, January 2019. ISSN 1572-9249.
8. Thomas A. Gerds, Per K. Andersen, and Michael W. Kattan. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191– 3203, 2014. ISSN 1097-0258.
9. Peter C. Austin, Frank E. Harrell, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Statistics in Medicine*, 39 (21):2714–2742, 2020. ISSN 1097-0258.
10. Adam R. Brentnall and Jack Cuzick. Risk Models for Breast Cancer and Their Validation. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 35(1):14–30, March 2020. ISSN 0883-4237.
11. Patrick Royston. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. *The Stata Journal*, 14(4):738–755, December 2014. ISSN 1536-867X.
12. Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics. Journal of the International Biometric Society*, pages 121–130, 1986.
13. D. R. Cox. Two Further Applications of a Model for Binary Regression. *Biometrika*, 45 (3/4):562–565, 1958. ISSN 0006-3444.
14. Marcel Wolbers, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, July 2014. ISSN 1465-4644.
15. Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, June 2013. ISSN 02776715.
16. Marcel Wolbers, Michael T. Koller, Jacqueline C. M. Witteman, and Ewout W. Steyerberg. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology*, 20(4):555–561, July 2009. ISSN 1044-3983.
17. Janez Stare, Maja Pohar Perme, and Robin Henderson. A Measure of Explained Variation for Event History Data. *Biometrics*, 67(3):750–759, 2011. ISSN 1541-0420.
18. Paul Blanche, Jean-Franc¸ois Dartigues, and H´el`ene Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, December 2013. ISSN 02776715.
19. Patrick J. Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, March 2005. ISSN 0006-341X, 1541-0420.
20. P. Saha and P. J. Heagerty. Time-Dependent Predictive Accuracy in the Presence of Competing Risks. *Biometrics*, 66(4):999–1011, December 2010. ISSN 0006341X.
21. Hans van Houwelingen and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, zeroth edition, November 2011. ISBN 978-0-429-09433-0.
22. N. van Geloven, Y. He, A. H. Zwinderman, and H. Putter. Estimation of incident dynamic AUC in practice. *Computational Statistics & Data Analysis*, 154:107095, February 2021. ISSN 0167-9473.
23. Giuliana Cortese, Thomas A. Gerds, and Per K. Andersen. Comparing predictions among competing risks models with time-dependent covariates. *Statistics in Medicine*, 32(18): 3089–3101, 2013. ISSN 1097-0258.
24. Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, February 2011. ISSN 03233847.
25. Michael W. Kattan and Thomas A. Gerds. The index of prediction accuracy: An intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research*, 2(1):7, December 2018. ISSN 2397-7523.
26. Andrew J Vickers, Angel M Cronin, Elena B Elkin, and Mithat Gonen. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*, 8(1):53, December 2008. ISSN 1472-6947.

7

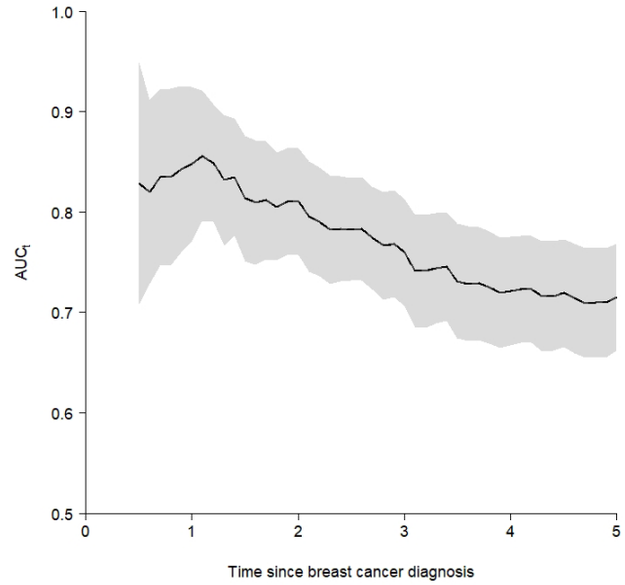
Supplementary material 5 - Supplementary Tables and Figures

Supplementary Table 1 Patient characteristics

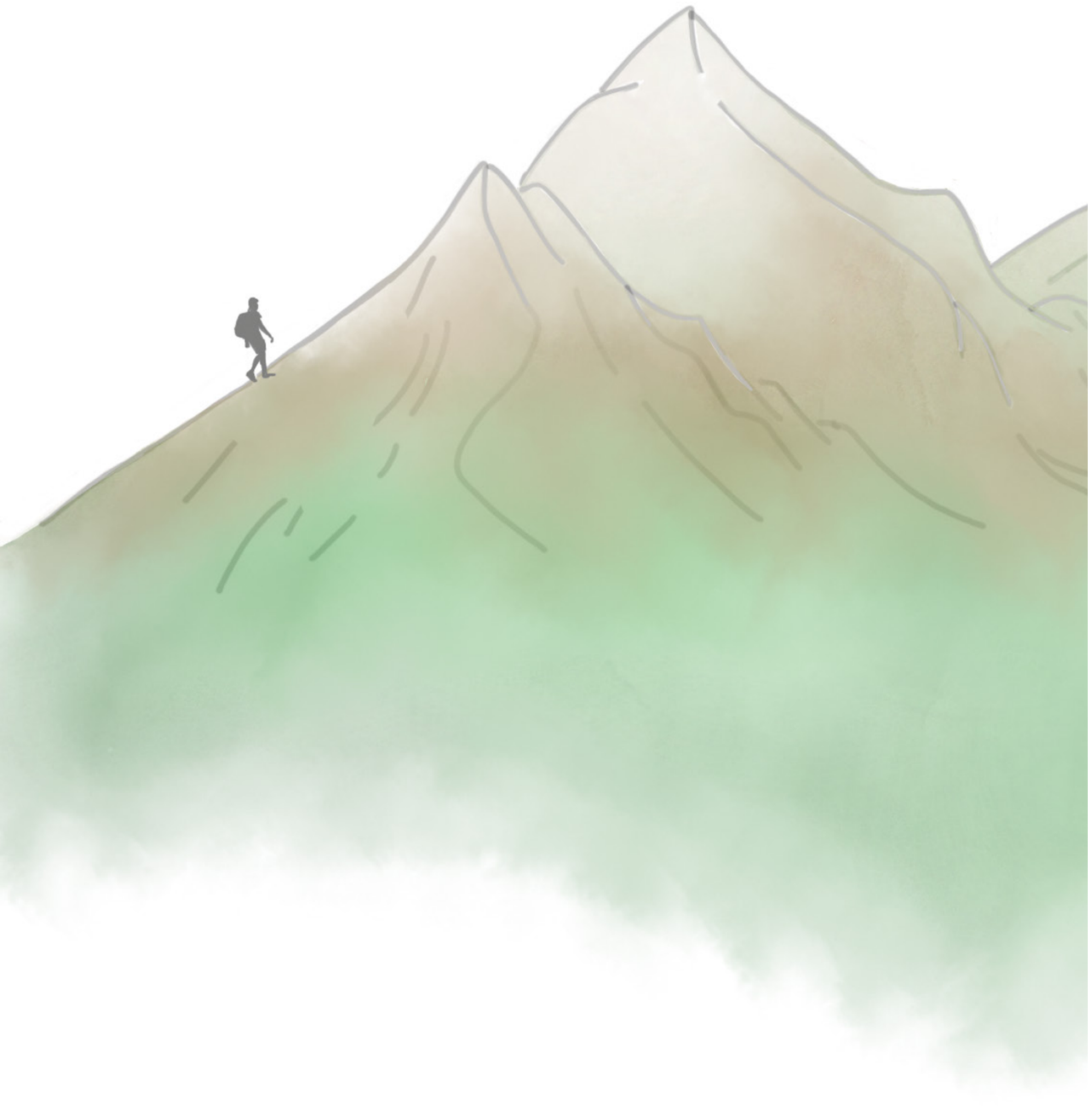
	Development cohort (N=1000)	Validation cohort (N=1000)
Age at diagnosis (years)		
Median [Min, Max]	74 [65, 95]	76.0 [70, 96]
Size of first tumour (cm)		
Median [Q1,Q3]	2.00 [1.40, 3.00]	1.80 [1.20, 2.60]
Nodal status (positive versus negative)		
Positive	358 (36%)	312 (31%)
Hormone Receptor status (ER+ and/or PR+ versus ER-/PR-)		
ER+ and/or PR+	822 (82%)	857 (86%)



Supplementary Figure 1: Cumulative incidence curves for breast cancer recurrence with death before recurrence as competing risk in the development (left) and validation (right) set. Dashed bars indicate 95% confidence intervals.



Supplementary Figure 2: Cumulative/dynamic time dependent AUC (AUCt) curve in the validation cohort. Time in years.



Chapter 8

General discussion

DISCUSSION

The aim of this thesis was to develop and validate a risk prediction model of contralateral breast cancer (CBC) for women with a first invasive breast cancer and to provide guidelines about performance assessment of risk prediction models with time-to-event outcomes. Large international population-based and hospital-based studies were analyzed to build the CBC prediction models, called PredictCBC models. In addition, the risk of CBC in women diagnosed with ductal carcinoma in situ (DCIS) was investigated. In this chapter, we will discuss the main findings and interpret them in a broader context with particular attention to the potential practical implications in clinical decision making. The methodological challenges of developing and validating a risk prediction model for time-to-event outcomes with and without competing risks are discussed in the CBC risk prediction context, and, more generally, using examples in the context of breast cancer recurrence and survival. Finally, future perspectives of research into prediction of CBC are also given. Future potential research directions in medical statistics are suggested based on methodological gaps that became apparent when analyzing real world breast cancer data.

Main findings and potential clinical implications

Risk factors and risk prediction of CBC in patients with first invasive breast cancer

Accurate CBC risk predictions are essential in clinical decision making regarding contralateral preventive mastectomies (CPMs) or other preventive strategies as personalized treatments and individualized surveillance. A number of patient-, first primary breast cancer-, and treatment characteristics have been suggested to be associated with CBC in several studies in the last 30 years[1-4]. In particular, patients' characteristics such as age at first breast cancer diagnosis, family history of breast cancer, hereditary mutations in the *BRCA1, BRCA2* and *CHEK2* (particularly the *c.1100delC* mutation) genes, as well as specific first primary breast cancer characteristics, i.e., tumor size, lymph node status and breast cancer histology, and (neo)adjuvant systemic therapies[5-11]. Furthermore, less attention has been given to study the association between CBC and lifestyle and reproductive factors. Currently, only a few factors such as body mass index (BMI) and parity (i.e., the number of births preceding first primary breast cancer diagnosis) have been suggested to be associated with CBC[12]. Could these characteristics (in statistical glossary labeled as predictors or covariates) be used to provide accurate CBC risk predictions and therewith support clinical decision making towards the choice of preventive strategies?

To answer this question, we developed and validated CBC risk prediction models (PredictCBC) using large population- and hospital-based studies mostly based in Europe, the United States and Australia with long follow-up (**Chapter 2**). The choice of the predictors in the analyses was based on evidence from the literature, availability of predictors in the studies and experience from clinical practice. PredictCBC models provided a moderate CBC prediction accuracy in terms of discrimination and calibration. These results were in line with other prediction tools currently available for first primary breast cancer and, generally speaking, in the oncology field[13]. Clinical decision making about CPM may be improved using PredictCBC models compared to current clinical practice in the Netherlands, which is mostly based on carriership of germline mutations in the *BRCA1/2* genes. We showed an overlap in the magnitude of CBC risk between *BRCA1/2* mutation carriers and non-*BRCA1/2* carriers. In fact, CPM might not be the preferred choice even in some patients with a *BRCA1/2* germline mutation when other characteristics are highly favorable. On the other hand, some additional preventive strategies might be considered among non-*BRCA1/2* carriers with unfavorable characteristics. We conclude that clinical decision making about preventive strategies should not only be based on a germline mutation in *BRCA1/2* genes or a bilateral breast cancer family history, but the multifactorial context should be considered using, for example, PredictCBC models. However, although PredictCBC models may more objectively estimate CBC risk, the decision making strongly depends on what patients and physicians consider an acceptable risk according to the guidelines and on patients' personal preferences.

Previously, two other tools were proposed to predict CBC: the Manchester formula and CBCrisk[14,15]. The former is a heuristic formula based on a literature review, while the latter is a CBC risk prediction model that was developed and validated in the United States around the same time we developed and validated the PredictCBC models[15,16]. Therefore, we decided to investigate the prediction accuracy of these two tools in comparison with PredictCBC models using the large population- and hospital-based studies we used to develop PredictCBC models (**Chapter 3**). We found that all three CBC tools provided only moderate prediction accuracy. We also found considerable heterogeneity among studies[17].

Other breast cancer risk genes, beyond *BRCA1/2,* have also been shown to be associated with CBC risk[10,18-21]. Can some of these additional genetic markers improve CBC risk prediction and clinical decision making? In **chapter 4,** we updated the PredictCBC models to PredictCBC-2.0 models. We incorporated data on a specific rare mutation in the *CHEK2* gene (*c.1100delC*), a polygenic risk score (PRS) combining 313 common genetic variants associated with breast cancer, and the previously shown relevant factors body mass index and parity[10,12,22]. We showed that the overall CBC prediction accuracy did not substantially increase between PredictCBC and PredictCBC-2.0 models. On the other hand, we demonstrated that the PredictCBC-2.0 model including *CHEK2 c.1100delC* and PRS had higher net benefit compared to the previous PredictCBC models. In other words, clinical decision making might additionally improve and be better tailored

8

incorporating common and specific rare genetic variants associated with CBC, especially among patients with *BRCA1/2* germline mutations and non-*BRCA1/2* carriers.

CBC risk in patients diagnosed with ductal carcinoma in situ

One of the most active research lines among physicians studying ductal carcinoma in situ (DCIS, a potential precursor of cancer) is whether a patient diagnosed with DCIS may develop a subsequent ipsilateral invasive breast cancer in the future[23]. However, it may be equally important to estimate the risk to develop a CBC in patients diagnosed with DCIS. In that light, the comparison of this risk to that of patients diagnosed with a first invasive breast cancer is relevant to help understand the magnitude of risk, the etiology, and treatment strategies. We showed that the CBC risk is slightly higher in patients with DCIS compared to invasive breast cancer patients using a large population-based dataset from 1989 to 2017 of the Dutch Cancer Registry, covering all breast cancer patients diagnosed in the Netherlands (**Chapter 5**). Around five out of 100 patients with DCIS may develop a CBC compared to around four patients with invasive breast cancer within 10 years. We illustrated that this slightly higher CBC risk in DCIS patients might be largely explained by the fact that adjuvant systemic therapies are not currently considered in DCIS patients according to the current Dutch guidelines. This does not imply that physicians should start treating DCIS patients with adjuvant systemic therapy since CBC risk is low and side effects of adjuvant systemic therapies have been demonstrated[24,25]. However, especially in the United States, more DCIS patients ask to undergo a CPM as a consequence of CBC risk overestimation[25]. We concluded that accurate prediction may be also needed for DCIS patients to facilitate decision making about additional treatments or CPM (**Chapter 5**).

Assessing the performance of survival and competing risks prediction models

In **chapter 6 and 7,** we propose frameworks for performance evaluation of predictions and for clinical utility of survival and competing risks models to provide guidance in the context of the STRATOS (STRengthening Analytical Thinking for Observational Studies) international initiative[26]. The objective of STRATOS is to provide accessible and guidance in the design and analysis of observational studies, since the quality of parts of biomedical research urgently needs improvement [26]. The members of the STRATOS initiative are experienced statisticians with different expertise in different topic groups (TG) (https://www.stratos-initiative.org/groups). Two topic groups (TG6 and TG8) involve statisticians with expertise in prediction models and survival analysis, respectively. In collaboration with the members of these two TGs, we provide guidance for different traditional and novel measures that may be used to assess the performance of prediction for survival and competing risk models. Typically, specific time points (also defined time horizons) are chosen as relevant by physicians. For example, physicians may be interested in predicting CBC risk at 5 and 10 years since the first invasive breast cancer diagnosis (**chapter 2, 3 and 4**). For this purpose, several time-dependent discrimination and calibration performance measures were proposed in the literature[27,28]. In **chapter 6 and 7**, we briefly provided an overview of the measures currently available in the literature. Secondly, we suggested to prioritize some of the available measures to evaluate discrimination, calibration and clinical utility of survival and competing risk models according to the literature and software availability. We aimed to guide practitioners and researchers interested in prediction models providing data and the software code. In **chapter 6** we provided both R and SAS code to develop and validate a survival risk prediction model using two free available data sets: the German Breast Cancer Study Group and the Rotterdam breast cancer dataset[29,30]. In **chapter 7**, we used a random sample of 1,000 patients from the Female breast cancer in the elderly; Optimizing Clinical guidelines USing clinico-pathological & molecular Study (FOCUS) and from NCR data to develop and validate a risk prediction model of breast cancer recurrence in the presence of competing risk due to mortality[31]. Both R and SAS codes are freely available in GitHub repositories created and maintained by the author (https://github.com/danielegiardiello/Prediction_performance_survival and https://github.com/survival-lumc/ValidationCompRisks) to facilitate the connection between methodological developments and software availability.

Strengths and limitations of the data used

One of the most important strengths of the studies presented in this thesis is the use of large hospital- and population-based studies with follow-up information of all women diagnosed with invasive breast cancer between 1990 and 2017. Most of the studies were from the Netherlands: the Netherlands Cancer Registry (NCR) has good quality information about patients, first primary breast cancer characteristics, and treatments[32]. We were fortunate to include other studies from the Netherlands such as Amsterdam Breast Cancer Study (ABCS), Breast Cancer Outcome Study of Mutation carriers (BOSOM), Erasmus Medical Center (EMC) study, and Hereditary Breast and Ovarian cancer study (HEBON)[33,34]. Their contribution was essential to incorporate key information about germline mutations in the *BRCA1/2* genes and on the performance of CPM. Last but not least, we incorporated data from the Breast Cancer Association Consortium (BCAC) including studies from other European countries, the United States and Australia. Using BCAC, we included other potentially important information, i.e., the specific rare mutation in the *CHEK2* gene (*c.1100delC*) and the PRS which was developed using BCAC data in a previous study[35,36]. The studies included in this thesis, after combining and harmonizing different sources of data, comprised more than 100,000 women diagnosed with invasive breast cancer or DCIS.

One of the most challenging parts of using different studies was missing data. In **chapter 2, 3 and 4,** the NCR represented more than 60% of all data to develop and

validate PredictCBC and PredictCBC-2.0 models. Family history for breast cancer, germline genetic information (i.e., *BRCA1/2* germline mutation, *CHEK2* c.1100delC and PRS), and CPM are completely unavailable in NCR. However, complete breast cancer characteristics and treatment information available in NCR contributed to developing good performance imputation models based on the correlation matrix of the data[37]. In addition, the remaining predictors were quite complete: more than 70% of patients had at most one missing predictor. Follow-up information regarding some outcomes were incomplete in some studies. For example, in some studies included in **chapter 2, 3 and 4**, CBC and CPM outcome information was incomplete leading to an underestimation of the cumulative incidence. This challenge can only be solved in the future by improvement of data collection at the source (i.e., the original registry from which the studies acquired their data).

Methodological challenges

There are several challenges in methodological research of clinical prediction models when individual data from different studies are available. The most important challenges characterizing all applied works (**chapter 2, 3 and 4**) of this thesis were missing data, heterogeneity between studies, and time-to-event outcomes with a special attention to competing risks.

Missing data are unavoidable in medical research and researchers tend to include only complete information to perform the statistical analyses. However, excluding a large proportion of information will lead to biases[38]. Biases may be substantial or negligible according to the reasons why data are missing. A common classification of missing data is: missing completely at random, missing at random and missing not at random[39]. According to the type of missing data classification, different statistical methods are suggested. These methods, commonly defined as multiple imputation, replace missing values with imputed values. To allow uncertainty about the missing data, imputed values are generated multiple times to create several different plausible imputed data. These results of all imputed data are combined (using different methods) to fully consider uncertainty among imputations. Multiple imputation methods are typically suggested when a missing is at random. When missing is completely at random, complete case analysis is suggested. On the other hand, when the amount of missing values is high, multiple imputation is recommended to avoid data reduction. Additional statistical investigations (e.g. sensitivity analyses) are required when missing is not at random[39]. An overview including the definition of the different types of missing data classification with an example and the suggested statistical method to minimize biased point and variance estimates is shown in **Table 1**.

Table 1: summary of missing data definitions and proposed solutions

Missing data mechanism	Definition	Example	Suggested solution
Missing completely at random	No systematic differences between the missing and observed values	Missing values on a certain predictor occur randomly in the data	Complete case analysis*
Missing at random	Missing values can be explained by differences in observed data	Patients with missing values of *BRCA1/2* tend to be older and with less aggressive tumors.	Multiple imputation
Missing not at random	Differences between missing and observed values may be still present after considering observed data	Patients with missing BMI may tend to have too high or too low self-reported BMI.	Multiple imputation and sensitivity analyses

* multiple imputation may be preferred especially when the amount of missing values is high.

In the context of individual data from multiple studies, missing values may be also systematic. Systematic missing data occurs when a predictor is completely unavailable in one or more studies. For example, genetic information is systematically missing in registry-based data as NCR. In this case, missing data can be considered at random since the missing can be fully explained by the fact that some studies simply did not collect this information. A potential solution for systematic missing values is to use the variable identifying the study as covariate to improve substantially the imputation models. More sophisticated multiple imputation approaches were proposed in the literature using, for example, mixed-effects imputation models to better consider heterogeneity between studies and the hierarchical nature of data[40-42]. Less is known about how to include competing risks in multiple imputation when individual data from multiple studies are available and in presence of systematic missing values[43,44].

Between-study heterogeneity should be also adequately considered both in the analysis, development and in validation of a risk prediction model using multiple studies. The heterogeneity may refer to different baseline risk for different studies, different distribution of the predictors among studies and/or different methods to measure outcomes and predictors. A risk prediction model can be developed using one-stage or a two-stage approach[45]. In one-stage individual patient data, a single model is developed and typically mixed-effects multilevel regression is used to consider within and between studies heterogeneity[42,46-48]. In two-stage individual patient data analysis, in the first step simple regression models are performed by study. Secondly, the estimates are combined using meta-analytic methods[42,47]. Both approaches have advantages and challenges, although some simulations showed a fully specified one stage approach should be preferred, especially in presence of systematic missing data[42,45]. However, few examples and guidelines are available in the literature with time-to-event outcomes[49-53].

8

In case of survival analysis, stratified Cox regression models or flexible parametric survival regression proposed by Royston and Parmar models may be used to account for different baseline risks among studies in one stage individual patient data analysis[53-55]. More sophisticated survival regression models were proposed (e.g. frailty models) in the literature[50]. Currently, to the best of our knowledge, no clear guidelines are available to clarify how to analyze individual patient data using multiple studies in the presence of competing risks outcomes. One of the first questions is whether Fine and Gray or cause-specific hazards models should be used, especially when the aim is to develop and validate a risk prediction model to predict the absolute risk[56,57]. In **chapters 2 and 4,** we developed PredictCBC models with a one stage individual patient data analysis approach using a stratified Fine and Gray method after multiple imputation of missing values. More practical and methodological efforts in the context of competing risks might be useful to better consider heterogeneity in the imputation and analysis models. **Table 2** summarizes the approaches of analyzing individual patient data using multiple studies.

Table 2: a summary of approaches to analyze individual patient data from multiple studies

Approach	Definition	Pros'	Cons'	Potential future developments
One-stage	A single model is developed where heterogeneity should be considered	- Simple; - Consistent with imputation of systematic missing data	- Sophisticated models (e.g. stratification, mixed-models, frailty models) - Computationally demanding	- Clear guidelines about how to develop, validate a competing risks / dynamic prediction models
Two-stage	A single model is developed by study. Estimates are combined using meta-analytical approaches	- Reasonable to fully consider heterogeneity among studies	- In case of systematic missing values, study-specific estimates may be diluted with multiple imputation	

IMPLICATIONS AND FURTHER RESEARCH

Potential clinical implications of PredictCBC models

With the work described in this thesis we have tried to pave a way for more accurate and potentially clinically relevant CBC risk prediction through PredictCBC models. Currently, decision making about CBC preventive strategies is essentially based on *BRCA1/2* germline mutation and/or family history. This choice is still reasonable and practical, although additional clinical and genetic information can refine clinical decision making about CPM. No clear guidelines are currently available about risk management of CBC. For example, in the Netherlands, most of them are based on the risk management for a first primary breast cancer diagnosis[58]. In some circumstances, the BOADICEA risk prediction model, originally developed to predict the risk of first primary breast cancer, is used to have a better idea about the CBC risk[59,60]. However, the risk to develop a CBC is higher among first breast cancer patients compared to the risk to develop a first primary breast cancer among healthy women[61]. Furthermore, the current BOADICEA model does not include crucial additional information from the primary breast cancer, most important being systemic therapies. Last but not least, currently although CBC prediction tools are available like PredictCBC models and CBCrisk, these are not widely used in clinical practice[62]. Generally speaking, PredictCBC models and other CBC risk tools may be used to better identify high risk patients and reassure low risk patients who have worries and fears about the risk to develop a new primary tumor in the opposite breast. An appropriate use of the models may help patients, with a low predicted CBC risk, to avoid CPM or to opt for an alternative preventive strategy such as personalized screening programs.

Suggestions to promote the usage of PredictCBC models

Three fundamental points may encourage physicians to use PredictCBC models in clinical practice: model implementation, model validation and model updating. Model implementation is essential to support public health strategies. One of the most important goals of model implementation is to provide a user-friendly tool to efficiently communicate risks to patients. A well-implemented PredictCBC model may facilitate a more interactive discussion about CBC risk management between patients and physicians. In a parallel PhD trajectory based on the same project, we took a first step towards this implementation. In addition, misconception of the risk may be minimized since breast cancer patients generally tend to overestimate their CBC risk[63,64]. Online website and software applications are largely (and also freely) available nowadays to implement and periodically update risk prediction tools in practice, for example Shiny in R (https://shiny.rstudio.com/) and Evidencio (https://www.evidencio.com/home).

Model validation is important to evaluate a risk prediction in a setting different than the one used to develop the model. PredictCBC models were built using studies from Europe (most of them from the Netherlands), United States, and Australia. We strongly encourage to validate PredictCBC models especially in Asian and African studies to refine CBC risk prediction and to introduce new or updated CBC risk management guidelines in different countries.

PredictCBC models, and generally prediction models, should be periodically monitored over time to provide up-to-date prediction and performances. However, it is still challenging to establish when and how often a periodic surveillance of the risk prediction

performances should be provided, especially when healthcare policies are quite heterogeneous among countries and change over the time[65]. Continuous monitoring and updates of prediction models are expensive and computationally demanding. Centralization, harmonization, and standardization of health care data may represent one of the ongoing/new frontiers in (medical) statistics. Recent and future developments in the field of data engineering, data architecture, and data science may substantially accelerate the usage of more sophisticated electronic health records to answer etiological questions and to develop, validate and monitor risk prediction models.

Further research and future developments in CBC risk prediction

Risk prediction models may need updates and revision. PredictCBC models might be in the future updated as new predictors become available. There are still stimulating opportunities to improve CBC risk prediction performance. For example, polygenic risk scores based on common genetic variants may steadily improve as new biological insights become available. Although the 313-polygenic risk score and *CHEK2 c.1100delC* are currently unlikely to add substantial improvements of CBC risk prediction performance in the general population, better tailored clinical decision making for individual patients was certainly apparent in PredictCBC-2.0 models. Other germline variants in *CHEK2*, and also in the *ATM* and *PALB2* genes are suggested to be associated with higher first breast cancer and CBC risk[9,20,66]. Breast density is a well-established risk factor of first primary breast cancer and it has been suggested to be associated with an increased CBC risk[67,68]. BOADICEA and the CBC specific risk tool CBCrisk include breast density as a predictor; however, no clinical utility evaluation was provided yet. Further research is needed to investigate whether including information about *ATM*, *PALB2* and breast density may improve CBC prediction and decision making. All potential aforementioned predictors are measured at (around) the diagnosis of primary breast cancer. Breast density, lifestyle and reproductive factors may change over time and adequate statistical methods are needed to consider time-dependent predictors to estimate CBC risk over the time.

Conceptually, any kind of risk prediction is challenging, especially far away in the future. It is reasonable to think that risk predictions may improve as new and updated information becomes available close to the prediction time horizon. Breast cancer is a multi-state disease with clinically relevant intermediate outcomes such as, for example, recurrence (local, locoregional, distant) and CBC. Individual patient prognosis can really differ as the intermediate events occurs or information about modifiable risk factors (e.g., BMI or alcohol use) and biomarkers change after the first event (e.g., diagnosis of DCIS or first primary invasive breast cancer). Recently new methods have been developed to simultaneously model intermediate states and incorporate longitudinal data[69-74]. The multi-state and dynamic prediction modeling can be used to reveal the relations between different types of events and to estimate predictions. Prediction can be calculated considering patients' personal characteristics and clinical status prior to the time of prediction (time horizon) at specific landmark time after the first event (e.g., IBC or DCIS) or at the time of an intermediate event (e.g., CBC) occurs. We propose a graphical representation of a multi-state modelling for dynamic prediction, encapsulating the outline of this thesis, as a potential next step in breast cancer prediction. (**Figure 1**). Unfortunately, we were not able to implement any dynamic models because information of intermediate events was incomplete and most of the predictors were not collected over the follow-up time.

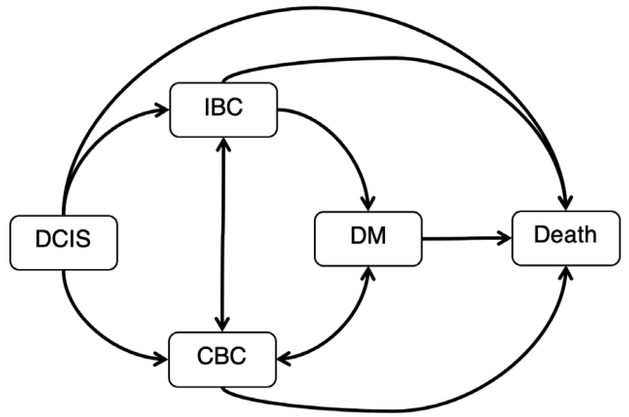


Figure 1: A simplified graphical representation of a multi-state modelling for breast cancer for dynamic prediction.
DCIS: ductal carcinoma in situ; IBC: ipsilateral breast cancer; CBC: contralateral breast cancer; DM: distant metastasis.

Perspectives and suggestions for further methodological research

As mentioned in the paragraphs before, no clear guidelines are currently available about how to analyze individual patient data with multiple studies in presence of competing risks. When missing values are present, how to incorporate competing risks outcomes in the imputation models is still an ongoing research, especially when multiple studies are included and systematic missing values can occur[41,43,44]. An overview about how to develop and validate a risk prediction model in presence of competing risks using individual data including multiple studies is really needed with a real application supported by software code for implementation. These potential guidelines and overviews might be extended in the context of multi-state and dynamic prediction modelling.

Cox proportional hazard models for each event and Fine and Gray regression are the

8

most known models for time-to-event data, although alternatives are possible[54,75]. Over the years, methodological research had focused on extending the potential violation of proportional hazard and random censoring assumptions. Other relevant issues include that many hospital- and population-based studies recruit patients at random times after diagnosis defined as left-truncation or delayed-entry. Few of the current discrimination measures (e.g. c-index and time-dependent Area Under the ROC curve) for time-to-event outcomes consider left-truncation[76]. Discrimination measures should be extended when left-truncation occurs. The potential violation of independent delayed entry assumption in parameter estimation and risk prediction might be additionally investigated. Simulation studies may be challenging in this setting[77-79].

Last but not least, prediction ignores technologies that will be discovered in the future[80]. There is an increasing interest in using and comparing machine learning and modern algorithms with standard statistical methods for risk prediction[81-83]. Thus, further comparison studies are welcome[84,85].

CONCLUSIONS

In conclusion, we paved the road for risk prediction of contralateral breast cancer (CBC) in patients diagnosed with first invasive breast cancer and ductal carcinoma in situ. The potential clinical utility and applications of CBC risk prediction models in clinical practice is largely investigated from a clinical viewpoint. An appropriate implementation and use of PredictCBC models may reassure patients about their fears to develop a new primary breast cancer in the opposite breast and to opt for alternative preventive strategies when their estimated CBC risk is low.

We sketched a framework for performance evaluation of predictions and clinical utility of survival and competing risks models using real word examples and providing the code of the statistical software currently used. There are still many stimulating and challenging opportunities to improve risk prediction and prognosis in the breast cancer field from genetics, biological and clinical perspectives. Much has been achieved in the last 30 years in medical statistics and biostatistics in risk prediction modelling, although identifying predictors that really improve prediction performances in (breast) cancer remains challenging. Further exciting opportunities lie ahead in methodological and applied research with the help of advance technological developments.

REFERENCES

1 Healey, E. A. *et al.* Contralateral breast cancer: clinical characteristics and impact on prognosis. *J Clin Oncol* **11**, 1545-1552, doi:10.1200/JCO.1993.11.8.1545 (1993).

2 Chen, Y., Thompson, W., Semenciw, R. & Mao, Y. Epidemiology of contralateral breast cancer. *Cancer Epidemiol Biomarkers Prev* **8**, 855-861 (1999).

3 Gao, X., Fisher, S. G. & Emami, B. Risk of second primary cancer in the contralateral breast in women treated for early-stage breast cancer: a population-based study. *Int J Radiat Oncol Biol Phys* **56**, 1038-1045, doi:10.1016/s0360-3016(03)00203-7 (2003).

4 Mariani, L. *et al.* Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Res Treat* **44**, 167-178, doi:10.1023/a:1005765403093 (1997).

5 van den Broek, A. J. *et al.* Impact of Age at Primary Breast Cancer on Contralateral Breast Cancer Risk in BRCA1/2 Mutation Carriers. *J Clin Oncol* **34**, 409-418, doi:10.1200/JCO.2015.62.3942 (2016).

6 Vichapat, V. *et al.* Risk factors for metachronous contralateral breast cancer suggest two aetiological pathways. *Eur J Cancer* **47**, 1919-1927, doi:10.1016/j.ejca.2011.05.004 (2011).

7 Vichapat, V. *et al.* Prognosis of metachronous contralateral breast cancer: importance of stage, age and interval time between the two diagnoses. *Breast Cancer Res Treat* **130**, 609-618, doi:10.1007/s10549-011-1618-8 (2011).

8 Reiner, A. S. *et al.* Hormone receptor status of a first primary breast cancer predicts contralateral breast cancer risk in the WECARE study population. *Breast Cancer Res* **19**, 83, doi:10.1186/s13058-017-0874-x (2017).

9 Reiner, A. S. *et al.* Breast Cancer Family History and Contralateral Breast Cancer Risk in Young Women: An Update From the Women's Environmental Cancer and Radiation Epidemiology Study. *J Clin Oncol* **36**, 1513-1520, doi:10.1200/JCO.2017.77.3424 (2018).

10 Weischer, M. *et al.* CHEK2*1100delC heterozygosity in women with breast cancer associated with early death, breast cancer-specific death, and increased risk of a second breast cancer. *J Clin Oncol* **30**, 4308-4316, doi:10.1200/JCO.2012.42.7336 (2012).

11 Akdeniz, D. *et al.* Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. *Breast* **44**, 1-14, doi:10.1016/j.breast.2018.11.005 (2019).

12 Akdeniz, D., Klaver, M. M., Smith, C. Z. A., Koppert, L. B. & Hooning, M. J. The impact of lifestyle and reproductive factors on the risk of a second new primary cancer in the contralateral breast: a systematic review and meta-analysis. *Cancer Causes Control* **31**, 403-416, doi:10.1007/s10552-020-01284-2 (2020).

13 Elmore, J. G. & Fletcher, S. W. The risk of cancer risk prediction: "What is my risk of getting breast cancer"? *J Natl Cancer Inst* **98**, 1673-1675, doi:10.1093/jnci/djj501 (2006).

14 Basu, N. N., Ross, G. L., Evans, D. G. & Barr, L. The Manchester guidelines for contralateral risk-reducing mastectomy. *World J Surg Oncol* **13**, 237, doi:10.1186/s12957-015-0638-y (2015).

15 Chowdhury, M., Euhus, D., Onega, T., Biswas, S. & Choudhary, P. K. A model for individualized risk prediction of contralateral breast cancer. *Breast Cancer Res Treat* **161**, 153-160, doi:10.1007/s10549-016-4039-x

(2017).
16 Chowdhury, M. *et al.* Validation of a personalized risk prediction model for contralateral breast cancer. *Breast Cancer Res Treat* **170**, 415-423, doi:10.1007/s10549-018-4763-5 (2018).
17 Giardiello, D. *et al.* Prediction of contralateral breast cancer: external validation of risk calculators in 20 international cohorts. *Breast Cancer Res Treat* **181**, 423-434, doi:10.1007/s10549-020-05611-8 (2020).
18 Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *J Mammary Gland Biol Neoplasia* **9**, 221-236, doi:10.1023/B:JOMG.0000048770.90334.3b (2004).
19 Kramer, I. *et al.* Breast Cancer Polygenic Risk Score and Contralateral Breast Cancer Risk. *Am J Hum Genet*, doi:10.1016/j.ajhg.2020.09.001 (2020).
20 Fanale, D. *et al.* Detection of Germline Mutations in a Cohort of 139 Patients with Bilateral Breast Cancer by Multi-Gene Panel Testing: Impact of Pathogenic Variants in Other Genes beyond BRCA1/2. *Cancers (Basel)* **12**, doi:10.3390/cancers12092415 (2020).
21 Mellemkjaer, L. *et al.* Risk for contralateral breast cancer among carriers of the CHEK2*1100delC mutation in the WECARE Study. *Br J Cancer* **98**, 728-733, doi:10.1038/sj.bjc.6604228 (2008).
22 Kramer, I. *et al.* Breast Cancer Polygenic Risk Score and Contralateral Breast Cancer Risk. *Am J Hum Genet* **107**, 837-848, doi:10.1016/j.ajhg.2020.09.001 (2020).
23 Visser, L. L. *et al.* Predictors of an Invasive Breast Cancer Recurrence after DCIS: A Systematic Review and Meta-analyses. *Cancer Epidemiol Biomarkers Prev* **28**, 835-845, doi:10.1158/1055-9965.EPI-18-0976 (2019).
24 Miller, M. E. *et al.* Contralateral Breast Cancer Risk in Women with Ductal Carcinoma In Situ: Is it High Enough to Justify Bilateral Mastectomy? *Ann Surg Oncol* **24**, 2889-2897, doi:10.1245/s10434-017-5931-2 (2017).
25 Tuttle, T. M. *et al.* Increasing rates of contralateral prophylactic mastectomy among patients with ductal carcinoma in situ. *J Clin Oncol* **27**, 1362-1367, doi:10.1200/JCO.2008.20.1681 (2009).
26 Sauerbrei, W. *et al.* STRengthening analytical thinking for observational studies: the STRATOS initiative. *Stat Med* **33**, 5413-5432, doi:10.1002/sim.6265 (2014).
27 Blanche, P., Dartigues, J. F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* **32**, 5381-5397, doi:10.1002/sim.5958 (2013).
28 Brentnall, A. R. & Cuzick, J. Risk Models for Breast Cancer and Their Validation. *Stat Sci* **35**, 14-30, doi:10.1214/19-STS729 (2020).
29 Schumacher, M. *et al.* Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *J Clin Oncol* **12**, 2086-2093, doi:10.1200/JCO.1994.12.10.2086 (1994).
30 Foekens, J. A. *et al.* The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Res* **60**, 636-643 (2000).
31 de Glas, N. A. *et al.* Postoperative complications and survival of elderly breast cancer patients: a FOCUS study analysis. *Breast Cancer Res Treat* **138**, 561-569, doi:10.1007/s10549-013-2462-9 (2013).
32 Font-Gonzalez, A. *et al.* Inferior survival for young patients with contralateral compared to unilateral breast cancer: a nationwide population-based study in the Netherlands. *Breast Cancer Res Treat* **139**, 811-819, doi:10.1007/s10549-013-2588-9 (2013).
33 Schmidt, M. K. *et al.* Breast Cancer Survival of BRCA1/BRCA2 Mutation Carriers in a Hospital-Based Cohort of Young Women. *J Natl Cancer Inst* **109**, doi:10.1093/jnci/djw329 (2017).
34 Pijpe, A. *et al.* Physical activity and the risk of breast cancer in BRCA1/2 mutation carriers. *Breast Cancer Res Treat* **120**, 235-244, doi:10.1007/s10549-009-0476-0 (2010).
35 Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94, doi:10.1038/nature24284 (2017).
36 Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21-34, doi:10.1016/j.ajhg.2018.11.002 (2019).
37 Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol* **110**, 63-73, doi:10.1016/j.jclinepi.2019.02.016 (2019).
38 Sterne, J. A. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**, b2393, doi:10.1136/bmj.b2393 (2009).
39 Van Buuren, S. *Flexible imputation of missing data*. Second edn, (Chapman and Hall/CRC, 2018).
40 Jolani, S., Debray, T. P., Koffijberg, H., van Buuren, S. & Moons, K. G. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med* **34**, 1841-1863, doi:10.1002/sim.6451 (2015).
41 Resche-Rigon, M. & White, I. R. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*, doi:10.1177/0962280216666564 (2016).
42 Kontopantelis, E. A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study. *Res Synth Methods* **9**, 417-430, doi:10.1002/jrsm.1303 (2018).
43 White, I. R. & Royston, P. Imputing missing covariate values for the Cox model. *Stat Med* **28**, 1982-1998, doi:10.1002/sim.3618 (2009).
44 Bartlett, J. W. & Taylor, J. M. Missing covariates in competing risks analysis. *Biostatistics* **17**, 751-763, doi:10.1093/biostatistics/kxw019 (2016).
45 Debray, T. P., Moons, K. G., Abo-Zaid, G. M., Koffijberg, H. & Riley, R. D. Individual participant data meta-analysis for a binary outcome: one-stage or two-stage? *PLoS One* **8**, e60650, doi:10.1371/journal.pone.0060650 (2013).
46 Debray, T. P., Koffijberg, H., Vergouwe, Y., Moons, K. G. & Steyerberg, E. W. Aggregating published prediction models with individual participant data: a comparison of different approaches. *Stat Med* **31**, 2697-2712, doi:10.1002/sim.5412 (2012).
47 Burke, D. L., Ensor, J. & Riley, R. D. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med* **36**, 855-875, doi:10.1002/sim.7141 (2017).
48 Ahmed, I., Debray, T. P., Moons, K. G. & Riley, R. D. Developing and validating risk prediction models in an individual participant data meta-analysis. *BMC Med Res Methodol* **14**, 3, doi:10.1186/1471-2288-14-3 (2014).
49 Steyerberg, E. W., Nieboer, D., Debray, T. P. A. & van Houwelingen, H. C. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. *Stat Med* **38**, 4290-4309, doi:10.1002/sim.8296 (2019).
50 Michiels, S., Baujat, B., Mahe, C., Sargent, D. J. & Pignon, J. P. Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses. *J Clin Epidemiol* **58**, 238-245, doi:10.1016/j.jclinepi.2004.08.013 (2005).

51 Bowden, J., Tierney, J. F., Simmonds, M., Copas, A. J. & Higgins, J. P. Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model. *Res Synth Methods* **2**, 150-162, doi:10.1002/jrsm.45 (2011).

52 Smith, C. T., Williamson, P. R. & Marson, A. G. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Stat Med* **24**, 1307-1319, doi:10.1002/sim.2050 (2005).

53 Westeneng, H. J. *et al.* Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *Lancet Neurol*, doi:10.1016/S1474-4422(18)30089-9 (2018).

54 Royston, P. & Parmar, M. K. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* **21**, 2175-2197, doi:10.1002/sim.1203 (2002).

55 Ensor, J. *et al.* Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. *Stat Med*, doi:10.1002/sim.8959 (2021).

56 Wolbers, M. *et al.* Competing risks analyses: objectives and approaches. *Eur Heart J* **35**, 2936-2941, doi:10.1093/eurheartj/ehu131 (2014).

57 Meddis, A., Latouche, A., Zhou, B., Michiels, S. & Fine, J. Meta-analysis of clinical trials with competing time-to-event endpoints. *Biom J* **62**, 712-723, doi:10.1002/bimj.201900103 (2020).

58 *Federatie Medisch Specialisten*, <https://richtlijnendatabase.nl/richtlijn/borstkanker/algemeen.html> (

59 Antoniou, A. C., Pharoah, P. P., Smith, P. & Easton, D. F. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br J Cancer* **91**, 1580-1590, doi:10.1038/sj.bjc.6602175 (2004).

60 Antoniou, A. C. *et al.* The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer* **98**, 1457-1466, doi:10.1038/sj.bjc.6604305 (2008).

61 Brenner, D. J. Contralateral second breast cancers: prediction and prevention. *J Natl Cancer Inst* **102**, 444-445, doi:10.1093/jnci/djq058 (2010).

62 O'Donnell, M. Estimating Contralateral Breast Cancer Risk. *Current Breast Cancer Reports* **10**, 91-97 (2018).

63 Abbott, A. *et al.* Perceptions of contralateral breast cancer: an overestimation of risk. *Ann Surg Oncol* **18**, 3129-3136, doi:10.1245/s10434-011-1914-x (2011).

64 Portschy, P. R. *et al.* Perceptions of Contralateral Breast Cancer Risk: A Prospective, Longitudinal Study. *Ann Surg Oncol* **22**, 3846-3852, doi:10.1245/s10434-015-4442-2 (2015).

65 Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res* **5**, 1, doi:10.1186/s41512-020-00090-3 (2021).

66 Breast Cancer Association, C. *et al.* Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. *N Engl J Med* **384**, 428-439, doi:10.1056/NEJMoa1913948 (2021).

67 Knight, J. A. *et al.* The association of mammographic density with risk of contralateral breast cancer and change in density with treatment in the WECARE study. *Breast Cancer Res* **20**, 23, doi:10.1186/s13058-018-0948-4 (2018).

68 McCormack, V. A. & dos Santos Silva, I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev* **15**, 1159-1169, doi:10.1158/1055-9965.EPI-06-0034 (2006).

69 Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R. & van de Velde, C. J. Estimation and prediction in a multi-state model for breast cancer. *Biom J* **48**, 366-380 (2006).

70 van Houwelingen, H. C. & Putter, H. *Dynamic prediction in clinical survival analysis*. (CRC Press, 2011).

71 Rizopoulos, D. *Joint models for longitudinal and time-to-event data: with applications in R*. (CRC Press, 2012).

72 Fontein, D. B. *et al.* Dynamic prediction in breast cancer: proving feasibility in clinical practice using the TEAM trial. *Ann Oncol* **26**, 1254-1262, doi:10.1093/annonc/mdv146 (2015).

73 Geskus, R. B. *Data analysis with competing risk and intermediate states*. (CRC press, 2016).

74 Martin, G. P., Sperrin, M., Snell, K. I. E., Buchan, I. & Riley, R. D. Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. *Stat Med* **40**, 498-517, doi:10.1002/sim.8787 (2021).

75 Haller, B., Schmidt, G. & Ulm, K. Applying competing risks regression models: an overview. *Lifetime Data Anal* **19**, 33-58, doi:10.1007/s10985-012-9230-8 (2013).

76 Li, S. Estimating time-dependent ROC curves using data under prevalent sampling. *Stat Med* **36**, 1285-1301, doi:10.1002/sim.7184 (2017).

77 Tsai, W.-Y. Testing the assumption of independence of truncation time and failure tim. *Biometrika* **77**, 169-177 (1990).

78 Niels, K. & Moeschberger, M. in *Survival analysis: state of art* (Springer, 1992).

79 Azzato, E. M. *et al.* Prevalent cases in observational studies of cancer survival: do they bias hazard ratio estimates? *Br J Cancer* **100**, 1806-1811, doi:10.1038/sj.bjc.6605062 (2009).

80 Taleb, N. N. *The black swan: the impact of the highly improbable* (Penguin books, 2007).

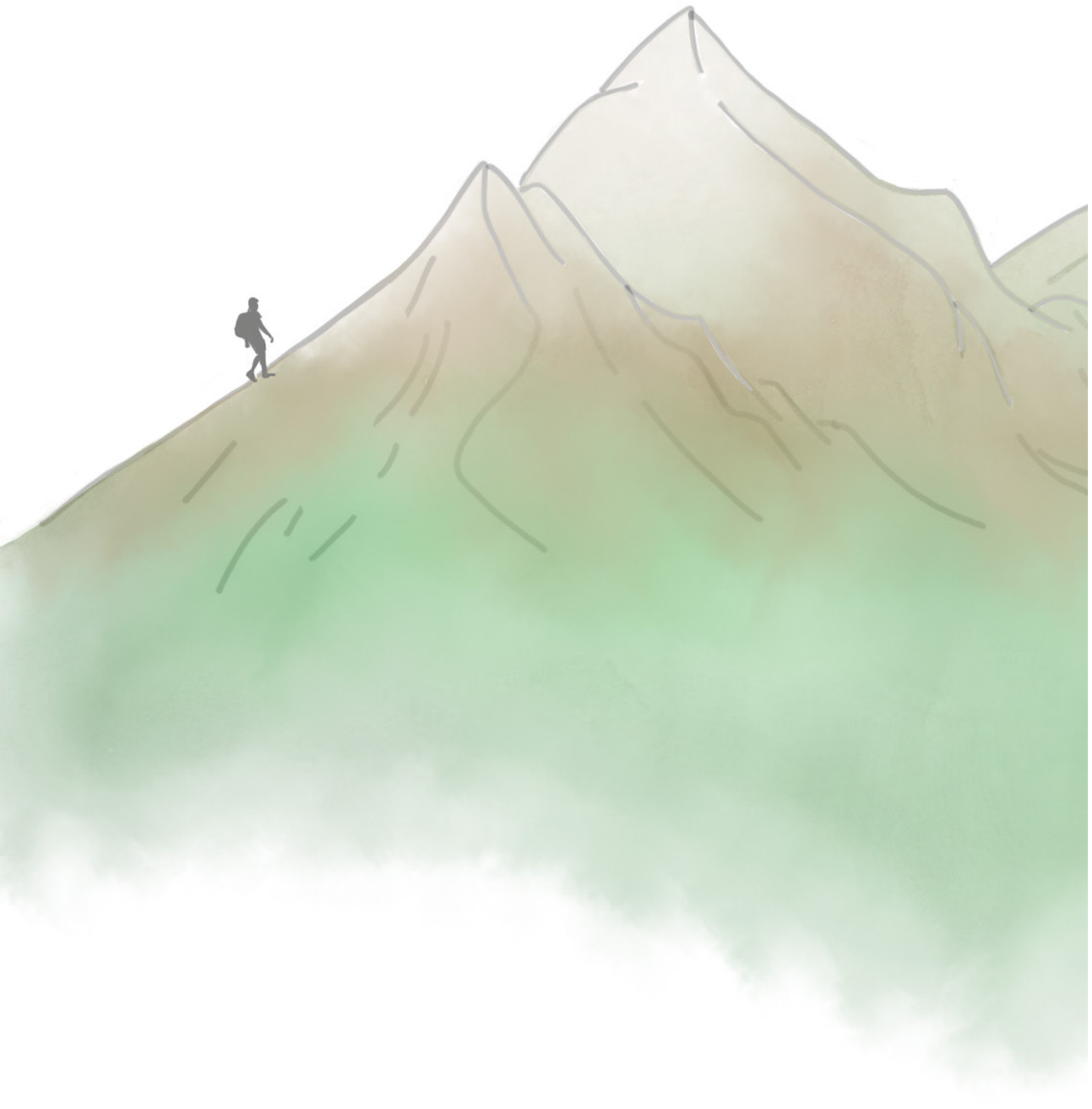
81 Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* **110**, 12-22, doi:10.1016/j.jclinepi.2019.02.004 (2019).

82 Wang, P., Li, Y. & Reddy, C. K. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys (CSUR)* **51**, 1-36, doi:10.1145/3214306 (2019).

83 Ming, C. *et al.* Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. *Breast Cancer Res* **21**, 75, doi:10.1186/s13058-019-1158-4 (2019).

84 Giardiello, D., Antoniou, A. C., Mariani, L., Easton, D. F. & Steyerberg, E. W. Letter to the editor: a response to Ming's study on machine learning techniques for personalized breast cancer risk prediction. *Breast Cancer Res* **22**, 17, doi:10.1186/s13058-020-1255-4 (2020).

85 Wilkinson, J. *et al.* Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* **2**, e677-e680, doi:10.1016/S2589-7500(20)30200-4 (2020).



Chapter 9

Summary

SUMMARY OF MAIN FINDINGS

Prediction of contralateral breast cancer

Breast cancer is the most common cancer among women worldwide. Although the incidence of breast cancer has increased, 10-year survival has improved approximately from 40% in 1960 and 1970 to almost 80% in 2010, in Europe. This rise may be attributed to early detection and better treatment modalities. The increase in diagnosis of first primary breast cancer implies that there are also more women at risk to develop a second primary tumor in the opposite breast. This is because contralateral breast cancer is the most common second primary cancer among women diagnosed with first breast cancer and accounts for approximately 40-50% of all new second cancers. On average, four to five out of 100 women with primary breast cancer develop a contralateral breast cancer within 10 years. These women have a worse prognosis compared with patients with unilateral breast cancer. Women with first primary breast cancer and high risk of contralateral breast cancer may opt for a contralateral preventive mastectomy, to almost nullify the risk to develop a contralateral breast cancer. In women at elevated contralateral breast cancer risk, such as women with pathogenic mutations in the *BRCA1*, *BRCA2*, or in the *CHEK2* genes or with a family history of (bilateral) breast cancer, the option of contralateral preventive mastectomy is actively discussed by clinicians. However, although contralateral preventive mastectomy is debatable in a large part of the breast cancer population without any genetic predisposition, an increasing number of women at low risk to develop a contralateral breast cancer choose to undergo a contralateral preventive mastectomy. Individualized contralateral breast cancer risk prediction may be potentially useful to facilitate shared decision making of physicians and patients regarding preventive strategies for those at high contralateral breast cancer risk and to avoid potentially unnecessary contralateral preventive mastectomies among patients at low risk to develop a contralateral breast cancer. One aim of this thesis was to develop and validate a contralateral breast cancer risk prediction model and evaluate its potential clinical utility (**chapter 1**).

In **chapter 2,** we developed and validated a contralateral breast cancer risk prediction model (PredictCBC). For this study, we used a large dataset of population- and hospital-based studies, mostly performed in Europe, United States and Australia, including more than 100,000 patients diagnosed with first invasive primary breast cancer between 1990-2013. PredictCBC provided the estimated 5- and 10-year risk to develop contralateral breast cancer using information about first primary breast cancer, family history, and *BRCA1/2* germline mutation status. We showed that the prediction performance accuracy of PredictCBC was moderate. PredictCBC may potentially tailor clinical decision making regarding preventive strategies that are currently essentially based on *BRCA1/2* germline mutation status. Contralateral preventive mastectomies might be unnecessary even in some patients with *BRCA1/2* germline mutations, especially among those with other favorable characteristics. On the other hand, preventive strategies such as personalized mammography screening might be necessary for non-*BRCA1/2* carrier patients, especially among those with unfavorable characteristics.

In **chapter 3,** we compared the prediction performance of PredictCBC with two other tools currently available to predict contralateral breast cancer: the Manchester formula and CBCrisk. The Manchester formula is a heuristic formula that estimates the lifetime contralateral breast cancer risk using information based on a systematic review of the literature. CBCrisk was developed using data on 1,921 contralateral breast cancer cases and 5,763 matched controls with first primary breast cancer. CBCrisk was externally validated in two independent studies in the United States. We externally validated the Manchester formula, and the contralateral breast cancer risk tools in the twenty studies used to develop and validate PredictCBC. We estimated that all three tools provided moderate individualized contralateral breast cancer prediction accuracy. For individual patients, we found a considerable heterogeneity of the prediction performances among the three tools. These differences reflect the heterogeneity among patients' characteristics and the corresponding contralateral breast cancer incidences among countries. We concluded that deeper biological and clinical insights, and the potential inclusion of genetic information beyond *BRCA1/2* germline mutation, might improve contralateral breast cancer prediction. In addition, this could further tailor clinical decision making about strategies for prevention or early detection of contralateral breast cancer. We encourage a more direct comparison between the three tools using large external datasets with complete information on all factors considered for contralateral breast cancer prediction models.

In **chapter 4**, we extended PredictCBC models by adding additional genetic information (e.g.: presence of the *CHEK2 c.11100delC* variant and a polygenic risk score based on 313 common genetic variants), and lifestyle and reproductive factors suggested to be associated with contralateral breast cancer. We developed and validated PredictCBC-2.0 models using updated follow-up information. We also extended the study population used to develop PredictCBC models including over 200,000 first primary breast cancer patients from a wide range of European-descendent studies diagnosed from 1990 to 2017. Additional genetic information beyond *BRCA1/2* germline mutation status improved contralateral breast cancer risk prediction. PredictCBC-2.0 might therefore help tailor clinical decision making towards contralateral preventive mastectomy or alternative preventive strategies such as personalized screening or personalized treatments.

In **chapter 5** we compared the contralateral breast cancer risk among patients diagnosed with first primary invasive breast cancer and patients diagnosed with ductal carcinoma

in situ, a potential precursor of cancer. We showed that the contralateral breast cancer risk is slightly higher in patients with ductal carcinoma in situ compared to invasive breast cancer patients using the Dutch cancer registry, a large population-based study in the Netherlands. Around five out of 100 patients with ductal carcinoma in situ may develop a contralateral breast cancer compared to around four patients among invasive breast cancer patients within 10 years. We concluded that this slightly higher contralateral breast cancer risk in ductal carcinoma in situ patients might be largely explained by that adjuvant systemic therapies in ductal carcinoma are not currently prescribed for ductal carcinoma in situ patients according to the current Dutch guidelines. This does not imply that we should start treating ductal carcinoma in situ patients with adjuvant systemic therapy since these patients have excellent prognosis and systemic therapy might have severe side effects. However, contralateral breast cancer prediction models may be useful for women with ductal carcinoma in situ as well to consider, for example, less or more intensified screening.

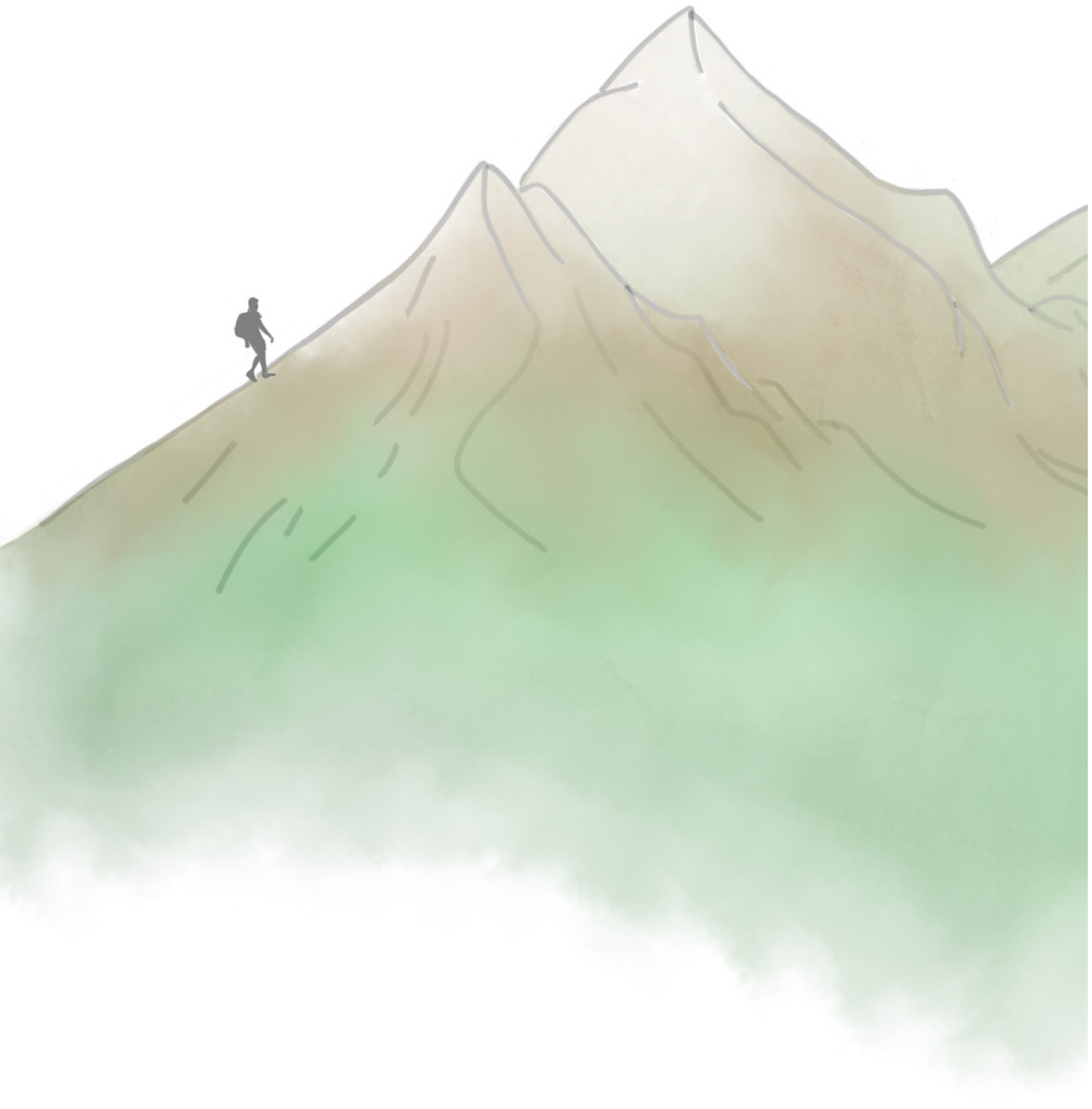
Assessing prediction performance with survival outcomes: practical guidance

Prediction research focuses on the development of well performing prediction models and on the assessment of their generalizability and applicability in clinical practice. A risk prediction model may be developed using regression, a statistical technique that estimates the relation between predictors and the outcome of interest (**chapter 1**). In many (breast) cancer studies, the outcome of interest is the time till an event occurs. Survival analysis is one of the most popular types of time-to-event analysis when the outcome is the survival time. When we study the occurrence of an event (e.g.: contralateral breast cancer) in a group of people, a person might not experience the event of interest over a certain time. In this case, survival time is censored. It might also happen that another event, different than the endpoint of interest, may preclude the event of interest from happening. For example, if we are studying contralateral breast cancer in women diagnosed with first primary breast cancer, some of them may die and those women will never be diagnosed with contralateral breast cancer since another competing (in this case fatal) event (or risk) occurred in their lives. The most common statistical regression models for survival analysis with or without competing risks are the Cox proportional hazard regression and the Fine and Gray regression model, respectively. These statistical regression models may be used to predict that an event of interest (e.g.: contralateral breast cancer) may occur in a certain time in the future (e.g.: within 10 years). Once a risk prediction model has been developed, it is first important to assess its performance. At the first instance, it is common to assess the prediction performance of a risk prediction in the same underlying population used to develop the model. This process is defined as internal validation. External validation refers to the evaluation of the prediction performance in a plausibly related population, which requires an independent dataset which may differ in setting, time, or place. If the prediction performance is sufficiently accurate, the risk prediction model might be applied in clinical practice to facilitate decision making (e.g.: by patient and physician considering contralateral preventive mastectomy). A further aim of this thesis was to provide guidance for assessing the prediction performance of time-to-event models with or without competing risks using motivating examples in breast cancer (**chapter 1**).

In **chapter 6**, we provided guidance and recommendations for assessing prediction performance of survival models. We described different measures that may be used to assess the performance of a prediction model with a survival outcome. We made a distinction between measures that can be used to assess the performance of predictions for specific time points (e.g., 5- or 10-year survival) and over a range of follow-up time. Prediction at specific time points will often be most relevant since clinicians and patients are usually interested in prognosis within a specified time. We illustrated how to develop a risk prediction model with survival outcome, how to assess its prediction performance and clinical utility through internal and external validation using real breast cancer datasets with the accompanying R and SAS software code.

In **chapter 7**, we provided an accessible overview of performance measures for a comprehensive assessment of the performance of a competing risks prediction model. We focused on how to validate a risk prediction model in the presence of competing risks at a given prediction horizon, a specified duration of time over which predictions are made (e.g., at 5 years). We extensively illustrated different methods on how to develop a risk prediction model with competing risks and how to calculate and interpret its prediction performance and clinical utility with illustration using a prediction model for breast cancer recurrence, including accompanying R code. Both overviews in **Chapter 6 and 7** were made on behalf of the international STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (http://stratos-initiative.org), which aims to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies for a non-specialist audience.

9



Appendices

NEDERLANDSE SAMENVATTING VAN DE BELANGRIJKSTE BEVINDINGEN

Het voorspellen van contralaterale borstkanker

Borstkanker is wereldwijd de meest voorkomende vorm van kanker onder vrouwen. Hoewel de incidentie van borstkanker is toegenomen, is de 10-jaars overleving in Europa verbeterd van ongeveer 40% in 1960 en 1970 tot bijna 80% in 2010. Deze verbetering kan worden toegeschreven aan vroegere opsporing en betere behandelingen. De toename in eerste primaire borstkankerdiagnoses impliceert dat er ook meer vrouwen het risico lopen op het ontwikkelen van een tweede primaire tumor in de andere borst. Dit is het geval omdat contralaterale borstkanker de meest voorkomende tweede primaire kanker is onder vrouwen gediagnosticeerd met een eerste borstkanker; ongeveer 40-50% van alle nieuwe tweede tumoren zijn contralaterale borsttumoren. Gemiddeld ontwikkelen vier tot vijf van de 100 vrouwen met primaire borstkanker contralaterale borstkanker binnen 10 jaar. Deze vrouwen hebben een slechtere prognose vergeleken met patiënten met unilaterale borstkanker. Vrouwen met een eerste primaire borstkanker en een hoog risico op contralaterale borstkanker kunnen kiezen voor een contralaterale preventieve mastectomie, met het doel om het risico op het ontwikkelen van contralaterale borstkanker zover mogelijk te beperken. In het geval dat vrouwen een verhoogd risico hebben op contralaterale borstkanker, zoals voor vrouwen met pathogene mutaties in de *BRCA1*- of *BRCA2*-genen of het *CHECK2*-gen, voor vrouwen met een familiegeschiedenis in (bilaterale) borstkanker, wordt de optie om voor een contralaterale preventieve mastectomie te kiezen actief besproken door clinici. Echter, hoewel contralaterale preventieve mastectomie onnodig conservatief is voor een groot deel van de borstkankerpopulatie zonder enige genetische aanleg, kiest een toenemend aantal vrouwen met een laag risico op contralaterale borstkanker voor een contralaterale preventieve mastectomie. Een geïndividualiseerde voorspelling van het risico op contralaterale borstkanker zou mogelijk nuttig kunnen zijn in de gezamenlijke besluitvorming door artsen en patiënten met betrekking tot preventieve strategieën voor vrouwen met een hoog risico op contralaterale borstkanker, en om mogelijk onnodige contralaterale preventieve mastectomieën onder patiënten met een laag risico te voorkomen. Eén van de doelen van dit proefschrift was om een voorspelmodel voor het risico op contralaterale borstkanker te ontwikkelen en te valideren en om de potentiele klinische bruikbaarheid van dit model te beoordelen (**hoofdstuk 1**).

In **hoofdstuk 2** ontwikkelden en valideerden we een voorspelmodel voor het risico op contralaterale borstkanker (PredictCBC). Voor dit onderzoek gebruikten we een grote database bestaande uit zowel studies onder algehele populaties als ziekenhuispopulaties, welke met name uitgevoerd waren in Europa, de Verenigde Staten en Australië, en in totaal meer dan 100.000 patiënten gediagnosticeerd met een eerste invasieve primaire borstkanker tussen 1990 en 2013 bevatte. PredictCBC schatte de 5- en 10-jaars risico's op het ontwikkelen van contralaterale borstkanker, met behulp van informatie over de eerste primaire borstkanker, familiegeschiedenis en *BRCA1/2*-kiembaanmutaties. We lieten zien dat de nauwkeurigheid waarmee PredictCBC voorspelde matig was. PredictCBC kan mogelijk artsen ondersteunen in het maken van beslissingen over preventieve strategieën, welke momenteel met name genomen worden op basis van de aanwezigheid van *BRCA1/2*-kiembaanmutaties. Contralaterale preventieve mastectomieën zouden zelfs onnodig kunnen zijn in bepaalde patiënten met *BRCA1/2*-kiembaanmutaties, vooral voor patiënten met andere gunstige karakteristieken. Aan de andere kant, preventieve strategieën zoals geïndividualiseerde mammografische screening zouden nodig kunnen zijn patiënten zonder *BRCA1/2*-kiembaanmutaties, met name voor diegenen met ongunstige karakteristieken.

In **hoofdstuk 3** vergeleken we de prestatie van het PredictCBC voorspelmodel met twee andere modellen die momenteel gebruikt worden voor het voorspellen van het risico op contralaterale borstkanker: de Manchester-formule en CBCrisk. De Manchester-formule is een methodische formule die het risico op contralaterale borstkanker gedurende de gehele levensduur schat, op basis van informatie uit een systematisch literatuuronderzoek. CBCrisk is ontwikkeld op basis van data van 1.921 patiënten met contralaterale borstkanker en 5.763 patiënten met een eerste primaire borstkanker. CBCrisk is extern gevalideerd in twee onafhankelijke studies in de Verenigde Staten. We valideerden de Manchester-formule en de risicomodellen voor contralaterale borstkanker extern in de 20 studies die gebruikt zijn voor het ontwikkelen en valideren van PredictCBC. We schatten dat alle drie de modellen matig accuraat waren in het voorspellen van het geïndividualiseerde risico op contralaterale borstkanker. Voor individuele patiënten vonden we aanzienlijke heterogeniteit in voorspelprestaties tussen de drie modellen. Deze verschillen weerspiegelen de heterogeniteit in patiëntkarakteristieken en de bijbehorende verschillen in incidentie van contralaterale borstkanker tussen landen. We concludeerden dat meer biologische en klinische inzichten en het mogelijk toevoegen van genetische varianten, naast de *BRCA1/2*-kiembaanmutaties, het voorspellen van het risico op contralaterale borstkanker zouden kunnen verbeteren. Daarnaast zou dit de klinische besluitvorming met betrekking tot preventie en vroege opsporing verder vorm kunnen geven. We moedigen een directere vergelijking tussen de drie modellen aan, met behulp van grote externe databases met complete informatie voor factoren die relevant zijn voor voorspelmodellen voor contralaterale borstkanker.

In **hoofdstuk 4** hebben we PredictCBC-modellen uitgebreid door extra genetische informatie (bijvoorbeeld de aanwezigheid van de *CHEK2 c.11100delC*-variant en een polygene risicoscore op basis van 313 veelvoorkomende genetische varianten), leefstijl

en reproductieve factoren toe te voegen. We ontwikkelden en valideerden PredictCBC-2.0-modellen met behulp van geupdate follow-upinformatie. We breidden daarnaast de studiepopulatie die gebruikt was voor het ontwikkelen van de PredictCBC-modellen uit, welke meer dan 200.000 patiënten met primaire borstkanker uit Europese studiepopulaties bevatte (diagnose tussen 1990 en 2017). Het toevoegen van extra genetische informatie, naast de *BRCA1/2*-kiembaanmutaties, verbeterde het voorspellen van het risico op contralaterale borstkanker. Het implementeren van PredictCBC-2.0 zou daarom kunnen bijdragen aan het verbeteren van klinische besluitvorming met betrekking tot preventieve mastectomie of alternatieve preventieve strategieën, zoals gepersonaliseerde screening of gepersonaliseerde behandelingen.

In **hoofdstuk 5** vergeleken we het risico op contralaterale borstkanker tussen patiënten gediagnosticeerd met een eerste primaire invasieve borstkanker en patiënten gediagnosticeerd met ductaal carcinoma in situ, een mogelijke voorloper van kanker. We toonden aan dat het risico op contralaterale borstkanker licht verhoogd is in patiënten met ductaal carcinoma in situ vergeleken met patiënten met invasieve borstkanker, door gebruik te maken van de Nederlandse Kankerregistratie, een landelijke databank met gegevens van kankerpatiënten in Nederland. Ongeveer vijf op de 100 patiënten met ductaal carcinoma in situ ontwikkelt binnen 10 jaar contralaterale borstkanker, tegen ongeveer vier patiënten onder 100 patiënten met invasieve borstkanker. We concludeerden dat dit licht verhoogde risico op contralaterale borstkanker mogelijk grotendeels veroorzaakt wordt doordat de huidige Nederlandse richtlijnen geen adjuvante systemische therapieën voorschrijven voor patiënten met ductaal carcinoma in situ. Dit betekent niet dat we patiënten met ductaal carcinoma in situ moeten gaan behandelen met adjuvante systemische therapieën, aangezien deze patiënten een excellente prognose hebben en systemische therapieën ernstige bijwerkingen kunnen hebben. Echter, voorspelmodellen voor contralaterale borstkanker kunnen ook nuttig zijn voor vrouwen met ductaal carcinoma in situ, bijvoorbeeld in het beslissen over de intensiteit van screening.

Het bepalen van de prestaties van voorspellingen met overlevingsuitkomsten: praktische handvatten

Prediction research richt zich op het ontwikkelen van goed functionerende voorspelmodellen en op het bepalen van hun generaliseerbaarheid en toepasbaarheid in de klinische praktijk. Een risicovoorspelmodel kan ontwikkeld worden door regressie-analyse te gebruiken, een statistische techniek die de relaties tussen voorspellers en de bestudeerde uitkomst kan inschatten (**hoofdstuk 1**). In veel (borst)kankerstudies is de bestudeerde uitkomst de tijd tot een *event* (een gebeurtenis zoals diagnose of studie-uitval) zich voordoet. *Survival analyse* is één van de meest populaire typen *time-to-event-analyses* als de uitkomst de periode van overleving is. Als we het vóórkomen van een gebeurtenis (bijvoorbeeld: contralaterale borstkanker diagnose) in een groep mensen bestuderen, dan kan het zijn dat een persoon de bestudeerde gebeurtenis in een bepaalde periode niet meemaakt. In dat geval wordt de informatie over overleving na die bepaalde periode niet meegenomen in de analyse (*censored survival time*). Het kan ook gebeuren dat een andere gebeurtenis, anders dan de bestudeerde uitkomst, uitsluit dat de bestudeerde uitkomst plaatsvindt. Bijvoorbeeld, als we geïnteresseerd zijn in het bestuderen van contralaterale borstkanker in een bepaalde groep vrouwen die gediagnosticeerd is met een eerste primaire borstkanker, dan zullen sommigen van hen overlijden. Deze vrouwen zullen nooit gediagnosticeerd worden met contralaterale borstkanker, omdat een ander, *competing event* (in dit geval een fatale), of *competing risk* in hun leven heeft plaatsgevonden. De meest gebruikte statistische regressiemodellen voor *survival analyse*, met of zonder *competing events*, zijn respectievelijk het *Cox proportional hazard* regressiemodel en het *Fine & Gray* regressiemodel. Deze statistische regressiemodellen kunnen gebruikt worden om te voorspellen of een bestudeerde gebeurtenis (bijvoorbeeld: contralaterale borstkanker) zal vóórkomen in een bepaalde periode in de toekomst (bijvoorbeeld: binnen 10 jaar). Als een risicovoorspelmodel ontwikkeld is, is het eerst belangrijk de prestatie te bepalen. In eerste instantie is het gebruikelijk om de voorspelprestatie te bepalen in de populatie die gebruikt is om het model te ontwikkelen. Dit proces wordt interne validatie genoemd. Externe validatie is het evalueren van de voorspelprestatie in een vergelijkbare onderzoekspopulatie, waarvoor een onafhankelijke dataset nodig is die kan verschillen wat betreft achtergrond, tijd, of plaats. Als de voorspelprestatie voldoende accuraat is, zou het risicovoorspelmodel toegepast kunnen worden in de klinische praktijk om klinische besluitvorming te faciliteren (bijvoorbeeld: door de patiënt en arts die contralaterale preventieve mastectomie overwegen). Een ander doel van dit proefschrift was om handvatten te bieden voor het bepalen van de voorspelprestatie van *time-to-event-modellen*, met of zonder *competing risks*, met behulp van voorbeelden toegepast op borstkanker (**hoofdstuk 1**).

In **hoofdstuk 6** bieden we handvatten en aanbevelingen voor het beoordelen van de voorspelprestatie van *survival modellen*. We beschreven verschillende maten die gebruikt kunnen worden om de prestatie van een voorspelmodel met overleving als uitkomst te kunnen bepalen. We maakten onderscheid tussen maten die gebruikt kunnen worden voor het bepalen van de prestatie van voorspellingen voor specifieke tijdstippen (bijvoorbeeld: 5- of 10-jaars overleving) en voorspellingen voor over de gehele periode waarvoor *time-to-event* gegevens beschikbaar zijn (*follow-up time*). Voorspellingen voor specifieke tijdstippen zullen vaak het meest relevant zijn, omdat artsen en patiënten meestal geïnteresseerd zijn in de prognose binnen een bepaalde tijdsperiode. We toonden aan hoe een risicovoorspelmodel met als uitkomst overleving te ontwikkelen, hoe zijn voorspelprestatie en klinische bruikbaarheid te bepalen door interne en

externe validatie, door gebruik te maken van bestaande borstkankerdatabases met de bijbehorende code in R en SAS.

In **hoofdstuk 7** boden we een toegankelijk overzicht van prestatiematen voor een uitgebreide beoordeling van de prestatie van een voorspelmodel met *competing risks*. We richtten ons op hoe we een risicovoorspelmodel konden valideren in de aanwezigheid van *competing risks* gegeven een bepaalde voorspelhorizon, een gespecificeerde tijdsduur waarover voorspellingen worden gemaakt (bijvoorbeeld: 5 jaar). We lieten uitgebreid verschillende methodes zien die gebruikt kunnen worden voor het ontwikkelen van risicovoorspelmodellen met concurrerende risico's en voor het berekenen en interpreteren van voorspelprestatie en klinische bruikbaarheid, geïllustreerd door een voorspelmodel voor de terugkeer van borstkanker, inclusief bijbehorende code in R. Beide overzichten in **hoofdstukken 6 en 7** werden gemaakt namens het internationale *STRengthening Analytical Thinking for Observational Studies (STRATOS)* initiatief (http://stratos-initiative.org), die tot doel heeft het bieden van toegankelijke en accurate richtlijnen voor onderwerpen die relevant zijn in het opzetten en analyseren van observationele studies voor een niet-gespecialiseerd publiek.

RIASSUNTO IN ITALIANO

Il tumore al seno o alla mammella (chiamato anche carcinoma mammario invasivo o infiltrante) è il tumore più frequente tra le donne di tutto il mondo. Questa patologia, sebbene diagnosticata ad un numero sempre crescente di donne, nel corso degli anni è diventata meno letale. Mentre nel 1960 la sopravvivenza a 10 anni dalla diagnosi era di circa 4 donne su 10, attualmente più di 8 donne su 10 riescono a sopravvivere in Europa a questa patologia. Questo sostanziale miglioramento è dovuto principalmente alla diagnosi precoce, resa possibile dagli esami mammografici periodici, e alle modalità di trattamento sempre più avanzate e personalizzate. Se da un lato l'anticipazione diagnostica e la sopravvivenza a questa malattia sono migliorate nel tempo, dall'altro un numero sempre maggiore di donne è potenzialmente a rischio di sviluppare un secondo tumore al seno opposto, conosciuto anche come tumore al seno controlaterale. Il tumore al seno controlaterale è un altro tumore primario (e non una recidiva del primo tumore alla mammella) ed è il più frequente secondo tumore che può essere diagnosticato tra le donne con un tumore alla mammella primario, rappresentando circa il 50% di tutti i secondi tumori. In media, circa 4 o 5 donne su 100 sviluppano un tumore anche nel rimanente seno sano entro 10 anni dalla diagnosi del primo, il che comporta una riduzione di durata della sopravvivenza rispetto alle donne con una diagnosi di un singolo tumore. Per questa ragione, donne con un tumore alla mammella primario o ad alto rischio di sviluppare un tumore alla mammella anche nel seno sano possono scegliere di effettuare la mastectomia preventiva controlaterale, ovverosia la rimozione del seno sano dopo che l'altro seno è stato colpito da un tumore. In donne con un elevato rischio di sviluppare un tumore al seno controlaterale, come quelle con una chiara predisposizione genetica alla malattia (ad esempio portatrici di una mutazione dei geni *BRCA1* e *BRCA2* o nel gene *CHEK2*), la mastectomia preventiva è una delle strategie preventive più considerate tra medici e pazienti. Nonostante l'efficacia della mastectomia preventiva non sia chiaramente provata soprattutto in assenza di predisposizioni genetiche, un numero crescente di donne a basso rischio di sviluppare un tumore al seno controlaterale decidono di sottoporsi alla mastectomia preventiva. Una stima del rischio oggettiva e basata sulle caratteristiche individuali delle pazienti può guidare la scelta riguardante le misure preventiva da adottare nella pratica clinica. Ad esempio, pazienti in cui viene riconosciuto un rischio elevato possono scegliere, dopo opportuna valutazione e discussione con una équipe di professionisti, di rimuovere completamente il seno sano attraverso la mastectomia, oppure possono optare di effettuare ulteriori trattamenti o screening mammari personalizzati. Al contrario, pazienti con basso rischio possono evitare di rimuovere il seno sano salvaguardandosi da eventuali effetti collaterali dovuti all'intervento chirurgico e potenziali impatti di natura psicologica. L'obiettivo principale di questa tesi di dottorato è stato quello di sviluppare un modello statistico che stimi il rischio di sviluppare un tumore al seno

controlaterale, di valutarne la sua capacità predittiva e di stabilire la sua potenziale utilità nel supportare e migliorare le scelte da intraprendere nella pratica clinica tra medico e paziente (**capitolo 1**).

Nel **capitolo 2** abbiamo sviluppato e validato un modello di rischio per il tumore al seno controlaterale, da noi denominato PredictCBC. I dati, di popolazione o ricavati da 20 studi clinici eseguiti principalmente in Europa, Stati Uniti d'America e Australia, provengono complessivamente da più di 100,000 donne con diagnosi di carcinoma mammario diagnosticato tra il 1990 e il 2013. Il modello PredictCBC fornisce una stima individualizzata del rischio di tumore al seno controlaterale a 5 e a 10 anni dalla diagnosi del carcinoma mammario primario, utilizzando informazioni individuali di ogni singola paziente quali: le caratteristiche del primo tumore al seno e dei corrispondenti trattamenti; la familiarità, intesa come una precedente diagnosi di tumore al seno tra i parenti di primo grado; la presenza di mutazione nei geni *BRCA1/2*. Abbiamo evidenziato come la mastectomia (preventiva) del seno sano potrebbe essere non giustificata perfino in alcune pazienti portatrici della mutazione nei geni *BRCA1/2*, specialmente tra coloro che hanno altre caratteristiche favorevoli. D'altro canto, misure che prevengano o possano diagnosticare precocemente il tumore al secondo seno come mammografie più frequenti e personalizzate potrebbero essere utili anche alle donne senza alcuna predisposizione genetica, qualora il rischio sia elevato per la presenza di altre caratteristiche sfavorevoli.

Nel **capitolo 3** abbiamo valutato la capacità previsionale del modello PredictCBC con quella di altri due modelli, denominati "la formula di Manchester" e il modello "CBCrisk", anch'essi sviluppati per stimare il rischio di tumore al seno controlaterale e attualmente disponibili in letteratura. La "formula di Manchester" è una formula euristica che stima il rischio di sviluppare il tumore nel corso della vita nel rimanente seno sano ed è stato messo a punto usando informazioni ricavate da una revisione sistematica precedentemente pubblicata. Lo strumento "CBCrisk", invece, è stato sviluppato utilizzando dei dati di studi provenienti da una casistica raccolta negli Stati Uniti. La comparazione fra questi tre strumenti è stata effettuata sui 20 studi precedentemente accennati in merito allo sviluppo del nostro modello. Come risultato, abbiamo verificato che mediamente tutti e tre gli strumenti forniscono una accuratezza previsiva moderata. Abbiamo inoltre osservato una considerabile eterogeneità delle prestazioni previsive tra i tre diversi strumenti in relazione alle differenze tra i vari studi per caratteristiche delle pazienti incluse e per *incidenza* di malattia. Da tali evidenze abbiamo dedotto che maggiori conoscenze biologiche, una migliore caratterizzazione clinica della malattia e più dettagliate informazioni genetiche potrebbero portare a prevedere meglio il rischio di tumore al seno controlaterale e conseguentemente a migliorare le decisioni cliniche. Anche la presenza di informazioni cliniche incomplete può compromettere la capacità previsionale dei modelli, ed è quindi auspicabile che in futuro la loro raccolta avvenga con maggiore accuratezza.

Nel **capitolo 4**, abbiamo esteso il modello PredictCBC sviluppato nel **capitolo 2** includendo ulteriori importanti informazioni associate alla patologia come informazioni relative allo stile di vita (ad esempio, l'indice di massa corporea), fattori riproduttivi (ad esempio, il numero totale di nascite dopo la gestazione), ed informazioni genetiche come la presenza della mutazione nel gene *CHEK2* e il valore del *rischio poligenico*. È stato sviluppato e validato un modello statistico più avanzato e aggiornato del precedente (denominato PredictCBC-2.0) utilizzando dati più aggiornati ed estesi che hanno incluso più di 200,000 pazienti. Abbiamo dimostrato che le ulteriori informazioni genetiche incluse nel nuovo modello forniscono una stima del rischio più accurata e potrebbero quindi migliorare la scelta delle strategie preventive attualmente previste.

Nel **capitolo 5**, abbiamo comparato il rischio di tumore al seno controlaterale tra le pazienti diagnosticate con il carcinoma mammario (invasivo o chiamato anche infiltrante) e il carcinoma duttale in situ. Quest'ultimo tipo di carcinoma è considerato una precancerosi, ovvero un fenomeno che potrebbe predisporre allo sviluppo di una forma invasiva. Utilizzando il registro tumori dei Paesi Bassi contenenti informazioni dettagliate riguardanti le pazienti, abbiamo stimato che il rischio di tumore al seno controlaterale è leggermente più alto tra le pazienti con il carcinoma duttale in situ rispetto alle pazienti con il carcinoma mammario invasivo. Questo leggero aumento del rischio di sviluppare un tumore al seno controlaterale tra le pazienti con il carcinoma duttale in situ potrebbe essere attribuito al fatto che nei Paesi Bassi queste pazienti non vengono trattate con terapie mirate (come la chemioterapia) che tendono in maniera primaria ad aumentare la probabilità di guarigione e a ridurre principalmente il rischio di potenziali recidive e metastasi, ma anche a contenere il rischio di un tumore al seno sano. Questo non implica che queste pazienti dovrebbero essere trattate con queste terapie mirate, soprattutto in virtu' del fatto che il carcinoma mammario in situ non rappresenta un rischio per la vita e le terapie potrebbero avere effetti collaterali che sopravanzano i benefici. Una più dettagliata e oggettiva quantificazione del rischio di tumore al seno controlaterale potrebbe essere utile anche per queste pazienti per poter stabilire strategie di prevenzione più mirate.

Guide pratiche per la valutazione delle prestazioni di previsione dei modelli di rischio di sopravvivenza

L'ambito della ricerca scientifica che si occupa della stima del rischio ha come obiettivo quello di sviluppare modelli previsivi accurati che possono essere applicabili nella pratica clinica. Un modello per la previsione del rischio può essere sviluppato utilizzando la regressione, una tecnica statistica che cerca di comprendere un fenomeno utilizzando

una serie di fattori (chiamati anche predittori). Ad esempio, se il fenomeno di interesse è la comparsa di una patologia (come, ad esempio, il tumore al seno controlaterale), si potrebbe essere interessati a comprendere come questa patologia sia legata all'età. In questo caso, l'occorrenza di tumore al seno controlaterale rappresenta il fenomeno di interesse mentre l'età è il predittore con il quale proviamo a comprendere meglio questo fenomeno. In molti studi sul cancro, come ad esempio gli studi sul tumore al seno, i ricercatori sono interessati a studiare il verificarsi di un determinato evento in un prefissato periodo di tempo. Nel caso del tumore al seno controlaterale, il fenomeno di interesse è il tempo (misurato in anni) trascorso dalla diagnosi del primo tumore al seno al secondo in un prefissato arco temporale (ad esempio a 5 o 10 anni dalla diagnosi del primo tumore). Questo fenomeno configura un capitolo della statistica definito come "analisi della sopravvivenza", per il quale sono stati predisposti appropriati strumenti di elaborazione. Secondo tale approccio, quando si studia un fenomeno che potrebbe accadere nel corso del tempo in un determinato gruppo di individui, per alcuni individui quel determinato evento potrebbe non accadere in un prefissato periodo di tempo. In questo caso il tempo è definito come tempo censorizzato. Un altro scenario possibile si verifica quando un altro fenomeno possa precludere il verificarsi dell'evento oggetto di studio. Per esempio, se si è interessati a studiare il verificarsi del tumore al seno in un prefissato periodo di tempo dopo la diagnosi del primo tumore, alcune pazienti potrebbero purtroppo non sopravvivere (per causa diretta o una causa non legata alla patologia di interesse). In questo caso, in queste pazienti non si potrà mai diagnosticare il tumore nel rimanente seno sano dato che un altro fenomeno (chiamato anche evento o rischio competitivo) si verifica durante il corso della loro vita. I più comuni modelli di regressione per l'analisi della sopravvivenza in assenza o in presenza di fenomeni/rischi/eventi competitivi sono, rispettivamente, il modello di regressione di Cox e il modello di Fine e Gray. Questi modelli possono essere utilizzati per prevedere il verificarsi di un determinato evento di interesse in un prefissato periodo di tempo. Per esempio, prevedere il rischio di tumore al seno controlaterale nei successivi 5 o 10 anni dalla diagnosi del primo tumore al seno. Una volta che un modello di previsione del rischio è stato sviluppato, è fondamentale valutarne la sua capacità e accuratezza previsiva. In primo luogo, la capacità previsiva di un modello viene valutata nella stessa popolazione utilizzata per sviluppare tale modello. Questo processo è definito come validazione o valutazione interna. La validazione esterna, invece, si riferisce alla valutazione della capacità e accuratezza di un modello previsivo in una popolazione diversa rispetto a quella per il quale il modello è stato sviluppato. La valutazione esterna permette di comprendere le ragioni per cui un modello è o non è generalizzabile in altre popolazioni o in altri scenari. Se la capacità previsiva di un modello fosse sufficientemente accurata, il modello previsivo potrebbe essere utile al fine di migliorare i processi decisionali nella pratica clinica (come, ad esempio, la scelta riguardante la mastectomia controlaterale preventiva). Un ulteriore obiettivo di questa tesi è stato di fornire delle guide pratiche per valutare la capacità previsiva di modelli di regressione per l'analisi della sopravvivenza in assenza o presenza di fenomeni/rischi/eventi competitivi utilizzando dati reali nell'ambito del tumore al seno.

Nel **capitolo 6**, abbiamo fornito una guida pratica che aiuti a comprendere come costruire un modello previsivo per fenomeni il cui studio rientra nell'analisi di sopravvivenza. Successivamente abbiamo descritto come valutare la capacità previsiva e come stimare la potenziale utilità di tali modelli nella pratica clinica, utilizzando dati riguardanti pazienti con tumore al seno, e abbiamo reso disponibile il corrispondente codice software per due linguaggi ampiamente utilizzati in ambito statistico, R e SAS.

Nel **capitolo 7**, abbiamo fornito una descrizione dettagliata e comprensibile delle attuali misure utilizzate per valutare la capacità previsiva di un modello di analisi della sopravvivenza in presenza di rischi competitivi. In particolare, abbiamo illustrato come sviluppare un modello per il calcolo del rischio di un fenomeno in presenza di eventi competitivi, come calcolare le diverse misure utili per valutarne capacità predittiva e utilità clinica, utilizzando dati reali di pazienti con tumore al seno e fornendo il corrispondente codice in R. Sia il **capitolo 6** che il **capitolo 7** sono stati scritti per conto di una iniziativa internazionale denominata STRATOS, che ha l'obiettivo di fornire guide e documentazioni accurate riguardanti argomenti di natura analitica ad un pubblico non specializzato.

Definizioni:

Incidenza: rappresenta la proporzione di individui che vengono colpiti da una determinata patologia in un determinato periodo di tempo. Misura il numero di nuovi casi nel periodo di tempo e individua il rischio (cioè la probabilità) che ha un individuo di contrarre una determinata patologia in quel periodo di tempo.

Rischio poligenico: è una misura che quantifica il potenziale rischio di sviluppare delle patologie basata sui geni e si calcola combinando gli effetti di un gran numero di varianti genetiche presenti nel genoma di ogni singolo individuo.

PUBLICATIONS

Giardiello D, Steyerberg EW, Hauptmann M, Adank MA, Akdeniz D, Blomqvist C, et al. Prediction and clinical utility of a contralateral breast cancer risk model. Breast Cancer Res. 2019 Dec 17;21(1):144.

Giardiello D, Hauptmann M, Steyerberg EW, Adank MA, Akdeniz D, Blom JC, et al. Prediction of contralateral breast cancer: external validation of risk calculators in 20 international cohorts. Breast Cancer Res Treat. 2020 Jun;181(2):423–34.

Giardiello* D, Kramer* I, Hooning MJ, Hauptmann M, Lips EH, Sawyer E, et al. Contralateral breast cancer risk in patients with ductal carcinoma in situ and invasive breast cancer. NPJ Breast Cancer. 2020 Nov 3;6(1):60.

Geloven N van, **Giardiello D**, Bonneville EF, Teece L, Ramspek CL, Smeden M van, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ. 2022 May 24;377:e069249.

** authors contributed equally*

Not in this thesis

Buisman FE, **Giardiello D**, Kemeny NE, Steyerberg EW, Höppener DJ, Galjart B, et al. Predicting 10-year survival after resection of colorectal liver metastases; an international study including biomarkers and perioperative treatment. Eur J Cancer. 2022 Jun 1;168:25–33.

Austin PC, Putter H, **Giardiello D**, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagn Progn Res. 2022 Dec;6(1):2.

van Seijen M, Lips EH, Fu L, **Giardiello D**, van Duijnhoven F, de Munck L, et al. Long-term risk of subsequent ipsilateral lesions after surgery with or without radiotherapy for ductal carcinoma in situ of the breast. Br J Cancer. 2021 Nov;125(10):1443–9.

van der Plas-Krijgsman* WG, **Giardiello* D**, Putter H, Steyerberg EW, Bastiaannet E, Stiggelbout AM, et al. Development and validation of the PORTRET tool to predict recurrence, overall survival, and other-cause mortality in older patients with breast cancer in the Netherlands: a population-based study. Lancet Healthy Longev. 2021 Nov 1;2(11):e704–11.

Huber V, Di Guardo L, Lalli L, **Giardiello D**, Cova A, Squarcina P, et al. Back to simplicity: a four-marker blood cell score to quantify prognostically relevant myeloid cells in melanoma patients. J Immunother Cancer. 2021 Feb;9(2):e001167.

Kramer I, Hooning MJ, Mavaddat N, Hauptmann M, Keeman R, Steyerberg EW, **Giardiello, D** et al. Breast Cancer Polygenic Risk Score and Contralateral Breast Cancer Risk. Am J Hum Genet. 2020 Nov 5;107(5):837–48.

Giardiello D, Antoniou AC, Mariani L, Easton DF, Steyerberg EW. Letter to the editor: a response to Ming's study on machine learning techniques for personalized breast cancer risk prediction. Breast Cancer Res. 2020 Feb 10;22(1):17.

Sobral-Leite M, Salomon I, Opdam M, Kruger DT, Beelen KJ, van der Noort V, **Giardiello, D** et al. Cancer-immune interactions in ER-positive breast cancers: PI3K pathway alterations and tumor-infiltrating lymphocytes. Breast Cancer Res. 2019 Aug 7;21(1):90.

Derks MGM, van de Velde CJH, **Giardiello D**, Seynaeve C, Putter H, Nortier JWR, et al. Impact of Comorbidities and Age on Cause-Specific Mortality in Postmenopausal Patients with Breast Cancer. Oncologist. 2019 Jul;24(7):e467–74.

Akdeniz D, Schmidt MK, Seynaeve CM, McCool D, **Giardiello D**, van den Broek AJ, et al. Risk factors for metachronous contralateral breast cancer: A systematic review and meta-analysis. Breast. 2019 Apr;44:1–14.

** authors contributed equally*

Separate from PhD trajectory

Necchi A, Sonpavde G, Lo Vullo S, **Giardiello D**, Bamias A, Crabb SJ, et al. Nomogram-based Prediction of Overall Survival in Patients with Metastatic Urothelial Carcinoma Receiving First-line Platinum-based Chemotherapy: Retrospective International Study of Invasive/Advanced Cancer of the Urothelium (RISC). Eur Urol. 2017 Feb;71(2):281–9.

Necchi A, **Giardiello D**, Mariani L. Methodological Considerations for Early-phase Development of Immune Checkpoint Inhibitors in Urothelial Bladder Cancer. Eur Urol. 2017 May;71(5):840–1.

Tuccitto A, Tazzari M, Beretta V, Rini F, Miranda C, Greco A, **Giardiello D** et al. Immunomodulatory Factors Control the Fate of Melanoma Tumor Initiating Cells. Stem Cells. 2016 Oct;34(10):2449–60.

Giacomini E, Ferrari N, Pitozzi A, Remistani M, **Giardiello D**, Maes D, et al. Dynamics of Mycoplasma hyopneumoniae seroconversion and infection in pigs in the three main production systems. Vet Res Commun. 2016 Jun;40(2):81–8.

Ferrari A, Lo Vullo S, **Giardiello D**, Veneroni L, Magni C, Clerici CA, et al. The Sooner the Better? How Symptom Interval Correlates With Outcome in Children and Adolescents With Solid Tumors: Regression Tree Analysis of the Findings of a Prospective Study. Pediatr Blood Cancer. 2016 Mar;63(3):479–85.

Raggi D, Mariani L, Giannatempo P, Lo Vullo S, **Giardiello D**, Nicolai N, et al. Prognostic reclassification of patients with intermediate-risk metastatic germ cell tumors: Implications for clinical practice, trial design, and molecular interrogation. Urol Oncol. 2015 Jul;33(7):332.e19-

Chiari M, Ferrari N, **Giardiello D**, Lanfranchi P, Zanoni M, Lavazza A, et al. Isolation and identification of Salmonella spp. from red foxes (Vulpes vulpes) and badgers (Meles meles) in northern Italy. Acta Vet Scand. 2014 Dec 10;56:86.

ACKNOWLEDGEMENTS

I would like to thank all people that they supported me during this journey. This work would have never been possible without your contribution.

I would like to express my gratitude to my promoter. Marjanka, thank you so much for giving me the opportunity to start this PhD journey, to sharpen my critical thinking, to learn about epidemiology of breast cancer and to have generally a wonderful experience in Amsterdam.

To my second promoter, Ewout, thank you for providing me the most updated literature about risk prediction modelling during the PhD, and for giving me the opportunity to collaborate with researchers inside and outside the Biomedical Data Sciences department in LUMC.

To my co-promoters Michael and Maartje thank you very much for your support.

Un grazie speciale a Luigi e al compianto Salvatore. Siete stati per i miei supervisori "sul campo" per tutta la prima durata del dottorato.

I would also like all (related) members in Marjanka's group and in the C2 department. It was wonderful to work with you and to share also "gezellige" moments. Special thanks to the "PhD room" – Maria, Iris, Anna. Felipe, Yuwei, Marcelo, Delal, Hayra for our discussions during our PhDs. Special thanks to Susanne, Renée, and Renske. In the era of personalized medicine, I will thank you individually. Special thanks to Maartje and Sina.

I am very grateful to my paranymphs: Yuwei and Matteo. Yuwei, thank you: I am lucky to have a good friend like you. Thank you for the nice talks and discussions about all kinds of topics.
Matteo, grazie per i tuoi consigli soprattutto nei miei momenti di difficoltà e per una serie di altre ragioni delle quali siamo entrambi consapevoli.

Cristian, thank you for your patience and to allow me to finish this work during my working time at EURAC.

I want to thank all patients who have their data available set for the studies in this thesis.
I want to thank all analysts, data managers, and physicians.
Grazie mamma, papà ed Elisa per concedermi la possibilità anche grazie ai vostri sacrifici di decidere con assoluta libertà tutte le scelte professionali che ho intrapreso fino ad

ora. Grazie per avermi fatto approdare in uno degli Olimpi della ricerca scientifica e per avermi sopportato durante i periodi più intensi.
Grazie Elisa e Stefano, ho una sorella e un cognato fantastici. Grazie per tutto quel che abbiamo condiviso durante la realizzazione di questo lavoro.

Un grazie infinito ai miei amici più stretti. Francesco, se le amicizie sono la famiglia che ti scegli, non potevo avere un fratello migliore con il quale condividere ogni aspetto delle nostre vite.
Grazie a Mattia, siamo la prova di come certi rapporti di amicizia possano fiorire nonostante una distanza di 1000 km. Attraverso te ringrazio di aver conosciuto altre belle persone come Andrea, Davide e Fabio.
Stefano, amico mio, da quel pomeriggio del 2008 all'Università siamo diventati due complici sia dal punto di vista professionale ma soprattutto personale. Grazie per ascoltarmi sempre: il tuo punto di vista e i tuoi valori per me sono e sempre saranno qualcosa di un valore inestimabile. Siamo diventati complementari e complici su un terreno di valori molto fertile.
Grazie Daniele, ai momenti condivisi insieme nonostante la distanza e soprattutto per i tuoi preziosi consigli.

Un ringraziamento speciale a tutte le persone che ho incontrato nei Paesi Bassi. A Massimo, per aver condiviso gli aspetti positivi e le criticità durante la permanenza ad Amsterdam e la tua ospitalità durante il mio periodo di transizione. A Gabriele e Giuliana per aver condiviso fatiche e soddisfazioni personali, professionali e sportive. Grazie a Irene, Stefano, Arianna.
Lief Rien, ik ben zo gelukkig om je te hebben ontmoeten. Bedankt voor je steun.
Thank you, John, for your practical support during this experience and for other personal reasons.

Grazie a coloro che durante questo percorso mi hanno inconsapevolmente aiutato anche con un semplice sorriso. Grazie a quelle persone che non conosco ma che sono per me fonte di ispirazione. Grazie a VIP Brianza: non so quanto durerà, siete il bellissimo dono.

ABOUT THE AUTHOR

Daniele Giardiello was born on November 28th, 1988, in Cantù (Como, Lombardy, Italy). In 2012 he obtained a Master of Science degree in Biostatistics and Experimental Statistics from the University of Milan-Bicocca (summa cum laude). In 2013, he moved to Brescia (Lombardy, Italy) to work as a statistician at the Zooprophylactic Institute of Lombardy and Emilia-Romagna to provide statistical support in veterinary research. From September 2014 to June 2016, he worked in Milan as a biostatistician at the National Cancer Institute under the supervision of dr. Luigi Mariani and under the daily mentoring of Salvatore Lo Vullo (in memoriam). Daniele was involved in multiple studies in oncology including melanoma, testicular cancer, pediatric cancer focusing on biostatistical models and machine learning methods especially for time-to-event outcomes. From June to September 2016, he worked as a statistician involved in design and analysis of phase II-III clinical trials at OPIS, a Contract Research Organization in Desio (Lombardy, Italy).

On November 8th 2016, Daniele moved to Amsterdam and started his PhD in the Division of Molecular Pathology of the Netherlands Cancer Institute – Antoni van Leeuwenhoek, and at the Department of Biomedical Data Sciences in Leiden University Medical Center under the supervision of prof. dr. ir. Marjanka Schmidt and prof. dr. Ewout Steyerberg. The results obtained during this academic research period (November 2016 – February 2021) are described in this thesis. Daniele collaborated and provided additional statistical support in many other projects inside and outside the departments. The scientific activity beyond his PhD trajectory is documented in the list of publications. He presented the results of his PhD at several international meetings and conferences

Since April 6th 2021, Daniele works as a biostatistician at EURAC, a private research center based in Bolzano/Bozen (South Tyrol, Italy) conducting research from a wide variety of scientific fields including biomedical research. He works at the Institute of Biomedicine under the supervision of dr. Cristian Pattaro.