FULL PAPER

BJC

British Journal of Cancer (2013) 109, 1296–1301 | doi: 10.1038/bjc.2013.382

Keywords: breast cancer incidence; validation; risk prediction model

Prospective validation of the breast cancer risk prediction model BOADICEA and a batch-mode version BOADICEACentre

R J MacInnis[1,2], A Bickerstaffe[2], C Apicella[2], G S Dite[2], J G Dowty[2], K Aujard[2], K-A Phillips[2,3,4,5], P Weideman[3], A Lee[6], MB Terry[7,8], G G Giles[1,2], M C Southey[9], A C Antoniou[6] and J L Hopper*[,2,10]

[1]Cancer Epidemiology Centre, Cancer Council Victoria, Victoria, Melbourne, Australia; [2]Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, The University of Melbourne, Victoria, Melbourne, Australia; [3]Division of Cancer Medicine, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia; [4]Department of Medicine, St Vincent's Hospital, The University of Melbourne, Victoria, Melbourne, Australia; [5]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Victoria, Melbourne, Australia; [6]Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK; [7]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA; [8]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA; [9]Department of Pathology, Genetic Epidemiology Laboratory, The University of Melbourne, Victoria, Melbourne, Australia and [10]School of Public Health, Seoul National University, Seoul, Korea

Background: Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) is a risk prediction algorithm that can be used to compute estimates of age-specific risk of breast cancer. It is uncertain whether BOADICEA performs adequately for populations outside the United Kingdom.

Methods: Using a batch mode version of BOADICEA that we developed (*BOADICEACentre*), we calculated the cumulative 10-year invasive breast cancer risk for 4176 Australian women of European ancestry unaffected at baseline from 1601 case and control families in the Australian Breast Cancer Family Registry. Based on 115 incident breast cancers, we investigated calibration, discrimination (using receiver-operating characteristic (ROC) curves) and accuracy at the individual level.

Results: The ratio of expected to observed number of breast cancers was 0.92 (95% confidence interval (CI) 0.76–1.10). The E/O ratios by subgroups of the participant's relationship to the index case and by the reported number of affected relatives ranged between 0.83 and 0.98 and all 95% CIs included 1.00. The area under the ROC curve was 0.70 (95% CI 0.66–0.75) and there was no evidence of systematic under- or over-dispersion ($P = 0.2$).

Conclusion: BOADICEA is well calibrated for Australian women, and had good discrimination and accuracy at the individual level.

Risk prediction models are important tools for identifying individuals at differing risks of developing a disease, especially if they take into account disease in relatives, and can be used to offer individually tailored prevention and clinical management. Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) is a risk prediction algorithm for breast and ovarian cancers that takes into account individual-specific data from relatives and computes *BRCA1* and *BRCA2* mutation carrier probabilities as well as age-specific risks for breast and ovarian cancer given a person's family history of cancer (Antoniou *et al*, 2004, 2008a). BOADICEA was developed using segregation analyses based on 2785 families ascertained through population-based studies of breast cancer and families with multiple affected individuals in which at least

*Correspondence: Professor JL Hopper; E-mail: j.hopper@unimelb.edu.au

Received 18 April 2013; revised 11 June 2013; accepted 22 June 2013; published online 13 August 2013

one family member had been screened for *BRCA1* and *BRCA2* mutations (Antoniou *et al*, 2008a). A web-interface allows users to easily compute carrier probabilities and future cancer risks (http://ccge.medschl.cam.ac.uk/boadicea/).

Independent validation of a risk model is important to justify implementation in clinical management. Although BOADICEA has been shown to be well calibrated in terms of predicting *BRCA1* and *BRCA2* mutation carrier status (Antoniou *et al*, 2008b; Stahlbom *et al*, 2012), it is still important to prospectively evaluate its performance in external populations, and in particular, whether BOADICEA performs adequately for populations outside the United Kingdom. Differences in underlying cancer incidences, founder mutations and environmental exposures could potentially lead to substantial under- or over-prediction of breast cancer risk for women living in different countries.

The aim of this study was to evaluate the performance of the BOADICEA model, in terms of calibration, validation and accuracy, in predicting first invasive breast cancer risks for female relatives of Australian women of European ancestry with (and without) breast cancer. In doing so, we introduce a tool for web-based risk calculation, *BOADICEACentre,* which can be run in batch mode.

PARTICIPANTS AND METHODS

Participants. Cohort participants were women who had not been diagnosed with invasive breast cancer at the time of recruitment to the Australian Breast Cancer Family Registry (ABCFR), who had completed a baseline questionnaire, and for whom follow-up data were available. This is an example of a prospective family study cohort (Hopper, 2011). The ABCFR is a component of the Breast Cancer Family Registry (John *et al*, 2004), and includes a population-based case–control-family study of the genetic, environmental and lifestyle factors associated with breast cancer; details of the recruitment strategy and baseline data collection methods have been previously described (Hopper *et al*, 1994, 1999; McCredie *et al*, 1998; Dite *et al*, 2003).

Population-based index cases were women under the age of 60 years when diagnosed with incident primary invasive breast cancer and were ascertained between 1992 and 1999 through population-based state cancer registries and recruited from metropolitan areas of Melbourne and Sydney, Australia (reporting of cancer to these state registries is a legislative requirement) (McCredie *et al*, 1998; John *et al*, 2004). Population-based index controls were also aged <60 years when identified from electoral rolls between 1992 and 1999. Family members of the index cases and controls were also recruited.

The ABCFR also includes 132 Melbourne-based families of Ashkenazi Jewish descent with one or more women with a history of breast cancer (Apicella *et al*, 2003), and 61 families with twin pairs in which one or both had a diagnosis of breast cancer recruited through the Australian Twin Registry (Hopper *et al*, 2013).

All participants provided written informed consent before participation and all studies were approved by the relevant local ethics committees.

After excluding women who had either a mastectomy or oophorectomy at baseline, the cohort consisted of 4176 women who were unaffected at baseline; 2704 from 991 population-based case families, 1322 (including index controls) from 521 population-based control families, 91 from 73 Ashkenazi Jewish families and 59 from 16 twin families.

Baseline questionnaire. At recruitment, all participants completed an interviewer-administered questionnaire which included detailed information on demographics, lifestyle and environmental factors, past surgeries (such as mastectomy and oophorectomy) and family history of cancer (see, McCredie *et al*, 1998; Dite *et al*, 2003; Milne *et al*, 2011).

Ascertainment of baseline family cancer history. At recruitment, all index cases completed a family history questionnaire (John *et al*, 2004) that sought cancer history for all their first-degree and second-degree relatives. Recruitment of parents, siblings and aunts of the index cases, and for some families other relatives depending on the cancer family history, was sought. Participating relatives then provided cancer history information on themselves and their relatives. A substantial proportion of the information collected for first-degree relatives was provided independently by the relatives themselves. Documented verification of reported cancers (through pathology reviews and reports, cancer registries and medical records) was sought wherever possible. Overall, 63% of breast cancers and 43% of non-breast cancers were verified.

Ascertainment of incident cancers. Incident breast cancer cases for participants were identified from notifications to the Victorian Cancer Registry and the NSW Central Cancer Registry of diagnoses of adenocarcinoma of the breast (International Classification of Diseases 9th revision rubric 174.0–174.9, or 10th revision rubric C50.0-C50.9) up to the end of 2010. Women with *in situ* breast cancers were not included as incident cases.

Imputation of missing family data. The data required from each of the participants and their relatives for these analyses were: relationship to the index case, date of birth, vital status, age at interview or death, and for those who had had cancer, the site and age at diagnosis. For some individuals (2% of the cohort), one or more of the above data items were missing and could not be calculated directly from known data. In these instances, data were imputed iteratively using a variation of a previously developed protocol (for more details, see Dite *et al*, 2003, 2010). Those with unknown age of breast cancer (4% of breast cancers reported) were assumed to have developed the disease at age 70 or last age of follow-up or age at death (if applicable), whichever age was the youngest.

***BRCA1* and *BRCA2* mutation testing.** Where available, information on *BRCA1* and *BRCA2* mutation testing was also taken into account. Mutations were protein-truncating or missense mutations classified as deleterious by the Breast Cancer Information Core (National Human Genome Research Institute, 2002). Details of testing are given elsewhere (Dite *et al*, 2010). *BRCA1* and *BRCA2* mutation testing of family members was conducted for 514 of 1857 prevalent cases (28%) identified at baseline, and for 79 of 4176 unaffected participants (2%). Sensitivity of the mutation detection technique was assumed to equal 70% and 80% for *BRCA1* and *BRCA2*, respectively.

BOADICEACentre. We have developed a set of new computer software programs written in Java, which we have named *BOADICEACentre.* These programs were used to format the pedigree files for BOADICEA, validate the data and to estimate missing values as required (see Appendix). *BOADICEACentre* was also used to automate the computing of risk estimates from BOADICEA (as the web version of BOADICEA only computes a risk estimate for one participant at a time) by automatically changing the participants, submitting pedigree files and collating the results.

Statistical analyses. For all participants, BOADICEA (University of Cambridge, Cambridge, UK) (Web program v3) was used to compute the risk of being diagnosed with invasive breast cancer during follow-up (Antoniou *et al*, 2004, 2008a). Follow-up began at baseline and ended at 10 years post baseline, age at death, age at first mastectomy (of any type), age at first oophorectomy (of any type) or at age 80 years, whichever came first. Australian and UK age-specific cancer incidences were used as the population reference incidences, based on the recent update of the BOADICEA model that incorporates calendar period cancer incidences up to 2010 and country-specific incidences. Risks were computed for women of Ashkenazi Jewish origin assuming *BRCA1* and *BRCA2* mutation prevalence for young controls (*BRCA1*: 1.6% and *BRCA2*: 1.2%; Satagopan *et al*, 2001; Antoniou *et al*, 2008a).

We evaluated the performance of BOADICEA by investigating multiple model properties including calibration, discrimination and accuracy at the individual level (or dispersion). Model calibration (which indicates the overall fit of the model) was evaluated by comparing the expected (E) number of cases computed from BOADICEA with the observed (O) number of breast cancer cases. To account for having multiple participants from the same family, we used a robust 95% confidence intervals (CI) for the ratio of expected to observed cases given by $\frac{E}{O} \pm 1.96\sqrt{Var\left(\log\left(\frac{E}{O}\right)\right)}$
where

$$Var\left(\log\left(\frac{E}{O}\right)\right) = \frac{\sum_{i=1}^{Nfam}\left(E_i - \frac{E}{O}O_i\right)^2}{\left(\sum_{i=1}^{Nfam} E_i\right)^2}$$

O = total observed cases, E = total expected, O_i = total observed within family i, E_i = total expected within family i. Summations are over all $Nfam$ families.

Discrimination (which is the ability of the model to distinguish between breast cancer cases and non-cases at the individual level) was evaluated using the area under the receiver-operating characteristic (ROC) curve, ignoring participants who were censored. For this measure, 0.5 is no better than chance and 1.0 is perfect discrimination. Sensitivity and specificity were assessed using cut-points of 10-year projected risks of 2%, 3%, 4% and 5%.

The accuracy of BOADICEA (i.e., whether the predicted probabilities fit the observed breast cancer status at the individual level) was evaluated using a logistic regression analysis in which the observed breast cancer status was the dependent variable and the log-odds of BOADICEA's predicted probability of developing breast cancer during the follow-up period was the independent variable. The null hypothesis, that the estimated regression coefficient was equal to 1 in the model without a constant term, was used to test for dispersion.

All other statistical analyses and graphs were performed using Stata 12.1 (Stata Corporation, College Station, TX, USA).

RESULTS

Cumulative 10-year invasive breast cancer risks were calculated for 4176 participants from 1601 families from the ABCFR, the characteristics of whom are shown in Table 1. Of these, 117 participants identified themselves as being of Ashkenazi descent, while 35 participants were known to be carriers of *BRCA1* or *BRCA2* mutations. Figure 1 shows the distribution of 10-year BOADICEA scores. During the 10 years of follow-up, a total of 115 incident invasive breast cancers were identified. Approximately 15% of the participants were censored before attaining 10 years of follow-up (average 5.7 years; 378 reached 80 years of age, 5 had a bilateral mastectomy, 232 had died because of causes other than breast cancer). Duration of follow-up of family members did not differ appreciably by status of the index case (average 9.2 years for case families and 9.6 years for control families).

The ratio of expected to observed number of breast cancers was 0.92 (95% CI 0.76–1.10; Table 2). There was some evidence that breast cancers were under-predicted for the 60–69 year age group (E/O ratio = 0.71, 95% CI 0.50–1.00). The E/O ratios by subgroups of the participant's relation to the index case and by the number of affected relatives reported ranged between 0.83 and 0.98 and all CIs included 1.00.

Table 1. Descriptive characteristics of the ABCFR cohort of 4176 women unaffected with breast cancer at baseline

	Unaffected after 10 years	Censored	Breast cancer within 10 years	Total
Overall	3452	609	115	4176
Age group				
20–29	621	19	2	642
30–39	856	27	16	899
40–49	755	44	27	826
50–59	663	55	31	749
60–69	516	86	32	634
70–79	41	378	7	426
Relationship to the index case–control				
Sister	1261	81	45	1387
Mother	347	179	22	548
Aunt	435	274	22	731
Other	1409	75	26	1510
Number of relatives with reported breast cancer				
0 First degree	1403	283	34	1720
1 First and 0 second degree	1186	162	35	1383
1 First and 1+ second degree	554	89	24	667
2+ First degree	309	75	22	406
Study type				
Families of index cases	2133	486	85	2704
Families of index controls/ Ashkenazis/twins	1319	123	30	1472

Abbreviation: ABCFR = Australian Breast Cancer Family Registry.

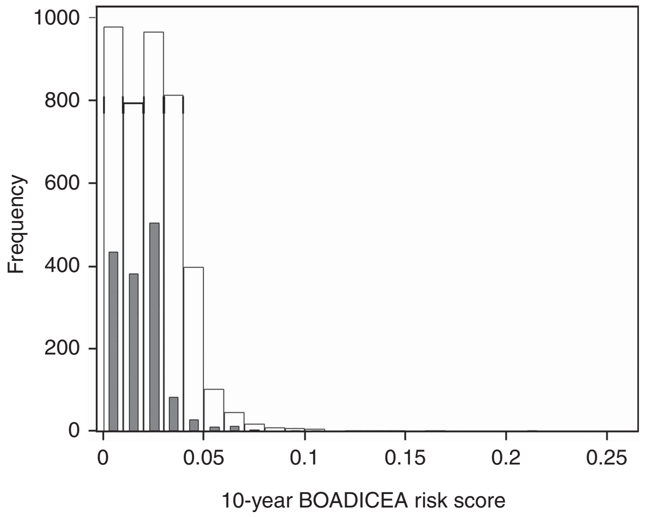


Figure 1. Distribution of BOADICEA scores. The white bars denote the distribution of 10-year BOADICEA risk scores for all participants; black bars denote the distribution for participants from index control families.

The E/O ratio point estimate was almost identical for 5-year follow-up (overall E/O = 0.93, 95% CI 0.71–1.21). Censoring

Table 2. Comparison of the observed number of incident breast cancers with the expected number based on BOADICEA, overall and by subgroups

	N	O	E	E/O	L95	U95
Overall	4176	115	105.6	0.92	0.76	1.10
Age group						
20–29	642	2	3.8	1.88	0.47	7.52
30–39	899	16	17.7	1.10	0.68	1.80
40–49	826	27	26.2	0.97	0.67	1.41
50–59	749	31	27.4	0.88	0.62	1.26
60–69	634	32	22.7	0.71	0.50	1.00
70–79	426	7	7.9	1.13	0.54	2.36
Relationship to the index case–control						
Sister	1387	45	41.0	0.91	0.68	1.22
Mother	548	22	19.0	0.86	0.57	1.31
Aunt	731	22	20.6	0.93	0.62	1.42
Other	1510	26	25.0	0.96	0.65	1.41
Number of relatives with reported breast cancer						
0 First degree	1720	34	32.1	0.94	0.67	1.32
1 First and 0 second degree	1383	35	34.4	0.98	0.71	1.37
1 First and 1 + second degree	667	24	20.8	0.87	0.58	1.29
2 + First degree	406	22	18.3	0.83	0.55	1.26
Study type						
Families of index cases	2704	85	77.4	0.91	0.74	1.13
Families of index controls/ Ashkenazi/twins	1472	30	28.2	0.94	0.66	1.34
Level of expected risk						
Quartile 1 (lowest)	1043	5	5.6	1.13	0.47	2.71
Quartile 2	1036	18	18.1	1.01	0.63	1.60
Quartile 3	1057	33	30.5	0.92	0.66	1.30
Quartile 4 (highest)	1040	59	51.4	0.87	0.67	1.12

Abbreviations: BOADICEA = Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm; E = total expected; O = total observed cases.

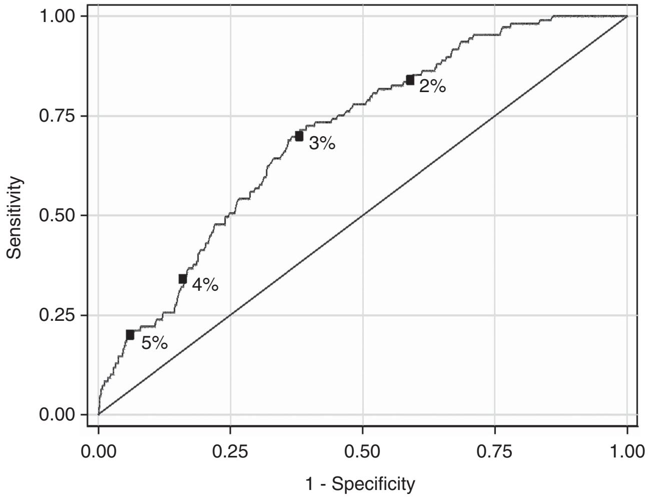


Figure 2. The receiver operator characteristic (ROC) curve BOADICEA as a predictor of 10-year cumulative risk. The square boxes on the ROC curve denote the cut-points. The area under the ROC curve was 0.70.

follow-up at the onset of any cancer (apart from non-melanoma skin cancer) did not materially alter the results (overall E/O = 0.91, 95% CI 0.75–1.10). Using UK incidences instead of Australian incidences made little difference to the results (e.g., overall E/O ratio = 0.93, 95% CI 0.78–1.12). Excluding families that had imputed data on one or more family members (161 families) also made virtually no difference to the results (data not shown).

Figure 2 shows that the test for discrimination (area under the ROC curve) was 0.70 (95% CI 0.66–0.75). The sensitivity and specificity given a 10-year BOADICEA risk score of 2%, 3%, 4% and 5% are highlighted in the figure. For example, sensitivity was 70% while specificity was 62% when a 10-year risk of 3% was used as a cut-point.

There was no evidence that the model was under- or over-dispersed (dispersion coefficient = 0.97, 95% CI 0.91–1.02, $P = 0.2$).

DISCUSSION

We have shown that BOADICEA is well calibrated for first invasive breast cancers overall and for most age and family history subgroups. In addition, the prediction model had good discriminatory accuracy, with no evidence of systematic under- or over-prediction at the individual level. We also provide details of a new BOADICEA utility, namely *BOADICEACentre,* which is easy-to-use and will help researchers with the input of data from an unlimited number of families in batch mode and the collation of results.

The under-prediction of BOADICEA in the 60–69 year old age group could be due to screening habits of women in the ABCFR. Wide-spread population screening in Australia has been reported for this age group (AIHW (Australian Institute of Health and Welfare), 2012), which may lead to over-diagnosis of breast cancer (Independent UK Panel on Breast Cancer Screening, 2012). Although the incidence rates that underlie BOADICEA should (at least partly) account for trends in population screening, it assumes that the sample being tested has similar characteristics to the overall population. The women in the ABCFR are, by study design, more likely to have a family history of breast cancer than the general population, and Australian women with a family history of breast cancer are more likely to seek additional screening than women without a family history (Roder *et al*, 2008). This discrepancy is likely to be most apparent for 60–69 year olds, an age range in which incidence peaks (Ferlay *et al*, 2010).

Use of Australian or UK incidences made little difference to the predictions. This is consistent with the fact that, over the past 30 years, the incidences for both countries were similar (Ferlay *et al*, 2010). Caution should be exercised in using BOADICEA for other populations in which incidences can vary markedly between and within countries, and between ethnic groups. A model based on an underlying population incidence helps take into account changing screening habits over time and differing lifestyle choices such as use of exogenous hormones. The recent extension of BOADICEA (Web program v3) allows users to select population-specific incidences.

The discriminatory power was good for BOADICEA, and compares favourably with other validated cancer risk prediction models. The Breast Cancer Risk Assessment Tool (BCRAT, also known as the Gail model; Gail *et al*, 1989) and the International Breast Cancer Intervention Study model (IBIS, also known as the Tyrer–Cuzick model; Tyrer *et al*, 2004) have been validated in different populations, with areas under the ROC curves ranging from 0.53 to 0.66 for BCRAT, and ∼0.70 for IBIS (Anothaisintawee *et al*, 2012; Quante *et al*, 2012). Amir *et al* (2003), using a small dataset from the United

Kingdom, reported slightly higher discriminatory power (area under the ROC curve ranging from 0.72 to 0.76) but poor calibration (E/O ranging from 0.48 to 0.81) for BCRAT, IBIS and BRCAPro. A Swedish study reported a slight underestimation (although not statistically significant) in the predicted number of invasive breast cancers (25 cases observed, E/O = 0.71, 95% CI 0.48–1.05) for an earlier version of BOADICEA (web program v1; Stahlbom *et al*, 2012).

Work is ongoing to include additional risk factors in BOADICEA to improve discriminatory accuracy, which will ultimately improve targeting of clinical interventions. The web-based version v3 now allows users to incorporate oestrogen and progesterone receptor, HER2, CK5/6 and CK14 status of diagnosed family members (Mavaddat *et al*, 2010). Ongoing extensions to BOADICEA include the incorporation of explicit other known breast cancer susceptibility variants, namely the associations of common alleles identified through genome-wide association studies and the effects of rare variants conferring moderate risks, as well as lifestyle/hormonal risk factors for breast cancer.

In summary, we have used prospective data and found that BOADICEA was well calibrated for a cohort of Australian women over-sampled for family history. These are the sorts of women seeking advice from clinicians about their risks of breast cancer, and those most likely to be referred to cancer family genetics services for genetic counselling. We are conducting similar prospective validation studies in other populations, and it will be important to compare the performance of BOADICEA with other breast cancer risk prediction models. Large data sets (potentially by combining data from several centres) will be needed to adequately assess the performance of BOADICEA for predicting ovarian and contralateral breast cancers. BOADICEA is freely and widely accessible through the internet and this study provides further support for its use by clinicians and women themselves.

ACKNOWLEDGEMENTS

The ABCFR is the Australian site of the Breast Cancer Family Registry, and this work was supported by grant UM1 CA164920 from the National Cancer Institute. The content of the manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centres in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products or organisations imply endorsement by the US Government or the BCFR. The Australian Breast Cancer Family Registry was also supported by the National Health and Medical Research Council (NHMRC) of Australia, the New South Wales Cancer Council and the Victorian Health Promotion Foundation. RJM is a NHMRC Sidney Sax Postdoctoral Research Fellow. K-AP is a National Breast Cancer Foundation Practitioner Fellow. ACA is a Cancer Research UK Senior Cancer Research Fellow (C12292/A11174). MCS is a NHMRC Senior Research Fellow and JLH is a NHMRC Senior Principal Research Fellow and was supported by the Distinguished Scholars Invitation Program at Seoul National University.

REFERENCES

AIHW (Australian Institute of Health and Welfare) (2012) *BreastScreen Australia monitoring report 2009-2010.* AIHW: Canberra.

Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, Wilson M, Howell A (2003) Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J Med Genet* **40**(11): 807–814.

Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A (2012) Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat* **133**(1): 1–10.

Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, Risch HA, Eyfjord JE, Hopper JL, Southey MC, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tryggvadottir L, Syrjakoski K, Kallioniemi OP, Eerola H, Nevanlinna H, Pharoah PD, Easton DF (2008a) The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer* **98**(8): 1457–1466.

Antoniou AC, Hardy R, Walker L, Evans DG, Shenton A, Eeles R, Shanley S, Pichert G, Izatt L, Rose S, Douglas F, Eccles D, Morrison PJ, Scott J, Zimmern RL, Easton DF, Pharoah PD (2008b) Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *J Med Genet* **45**(7): 425–431.

Antoniou AC, Pharoah PP, Smith P, Easton DF (2004) The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br J Cancer* **91**(8): 1580–1590.

Apicella C, Andrews L, Hodgson SV, Fisher SA, Lewis CM, Solomon E, Tucker K, Friedlander M, Bankier A, Southey MC, Venter DJ, Hopper JL (2003) Log odds of carrying an Ancestral Mutation in BRCA1 or BRCA2 for a Defined personal and family history in an Ashkenazi Jewish woman (LAMBDA). *Breast Cancer Res* **5**(6): R206–R216.

Dite GS, Jenkins MA, Southey MC, Hocking JS, Giles GG, McCredie MR, Venter DJ, Hopper JL (2003) Familial risks, early-onset breast cancer, and BRCA1 and BRCA2 germline mutations. *J Natl Cancer Inst* **95**(6): 448–557.

Dite GS, Whittemore AS, Knight JA, John EM, Milne RL, Andrulis IL, Southey MC, McCredie MR, Giles GG, Miron A, Phipps AI, West DW, Hopper JL (2010) Increased cancer risks for relatives of very early-onset breast cancer cases with and without BRCA1 and BRCA2 mutations. *Br J Cancer* **103**(7): 1103–1108.

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (2010) *GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10.* International Agency for Research on Cancer: Lyon, France.

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* **81**(24): 1879–1886.

Hopper JL (2011) Disease-specific prospective family study cohorts enriched for familial risk. *Epidemiol Perspect Innov* **8**(1): 2.

Hopper JL, Chenevix-Trench G, Jolley DJ, Dite GS, Jenkins MA, Venter DJ, McCredie MR, Giles GG (1999) Design and analysis issues in a population-based, case-control-family study of the genetic epidemiology of breast cancer and the Co-operative Family Registry for Breast Cancer Studies (CFRBCS). *J Natl Cancer Inst Monogr* **26**: 95–100.

Hopper JL, Foley DL, White PA, Pollaers V (2013) Australian Twin Registry: 30 years of progress. *Twin Res Hum Genet* **16**(1): 34–42.

Hopper JL, Giles GG, McCredie MRE, Boyle P (1994) Background, rationale and protocol for a case-control-family study of breast cancer. *The Breast* **3**: 79–86.

Independent UK Panel on Breast Cancer Screening (2012) The benefits and harms of breast cancer screening: an independent review. *Lancet* **380**(9855): 1778–1786.

John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT, Ziogas A, Andrulis IL, Anton-Culver H, Boyd N, Buys SS, Daly MB, O'Malley FP, Santella RM, Southey MC, Venne VL, Venter DJ, West DW, Whittemore AS, Seminara D (2004) The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res* **6**(4): R375–R389.

Mavaddat N, Rebbeck TR, Lakhani SR, Easton DF, Antoniou AC (2010) Incorporating tumour pathology information into breast cancer risk prediction algorithms. *Breast Cancer Res* **12**(3): R28.

McCredie MR, Dite GS, Giles GG, Hopper JL (1998) Breast cancer in Australian women under the age of 40. *Cancer Causes Control* **9**(2): 189–198.

Milne RL, John EM, Knight JA, Dite GS, Southey MC, Giles GG, Apicella C, West DW, Andrulis IL, Whittemore AS, Hopper JL (2011) The potential value of sibling controls compared with population controls for association studies of lifestyle-related risk factors: an example from the Breast Cancer Family Registry. *Int J Epidemiol* **40**(5): 1342–1354.

National Human Genome Research Institute (2002) Breast cancer information core: an open access on-line breast cancer mutation data base. (http://research.nhgri.nih.gov/bic/).

Quante AS, Whittemore AS, Shriver T, Strauch K, Terry MB (2012) Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to differential model performance. *Breast Cancer Res* **14**(6): R144.

Roder D, Houssami N, Farshid G, Gill G, Luke C, Downey P, Beckmann K, Iosifidis P, Grieve L, Williamson L (2008) Population screening and intensity of screening are associated with reduced breast cancer mortality: evidence of efficacy of mammography screening in Australia. *Breast Cancer Res Treat* **108**(3): 409–416.

Satagopan JM, Offit K, Foulkes W, Robson ME, Wacholder S, Eng CM, Karp SE, Begg CB (2001) The lifetime risks of breast cancer in Ashkenazi Jewish carriers of BRCA1 and BRCA2 mutations. *Cancer Epidemiol Biomarkers Prev* **10**(5): 467–473.

Stahlbom AK, Johansson H, Liljegren A, von Wachenfeldt A, Arver B (2012) Evaluation of the BOADICEA risk assessment model in women with a family history of breast cancer. *Fam Cancer* **11**(1): 33–40.

Tyrer J, Duffy SW, Cuzick J (2004) A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* **23**(7): 1111–1130.

APPENDIX A

Batch processing with the web-based BOADICEA program

A.1 Overview. The BOADICEA web client (http://ccge.medschl.cam.ac.uk/boadicea/) is an interactive tool for calculating risk estimates for breast cancer and *BRCA1* and *BRCA2* mutation status. The user enters a pedigree (manually or using file upload) and specifies an individual for risk estimation.

Researchers often need to calculate risk estimates from many pedigrees and for many individuals in each pedigree. The current BOADICEA web client makes this task laborious and time-consuming; each pedigree file must be free from errors, missing data may need to be imputed and risk estimates can only be requested for one member of one pedigree at a time.

To address these difficulties, we developed *BOADICEACentre*, an open source software tool written in Java that adopts a streamlined approach to batch processing. BOADICEACentre incorporates pedigree targeting, data validation, missing data imputation, online submission and collation of results for the web-based BOADICEA program.

A.2 Producing targeted pedigrees. BOADICEACentre produces a BOADICEA format pedigree file for each individual identified in a target list. The tool can also be configured to compute results for *every* woman in every pedigree.

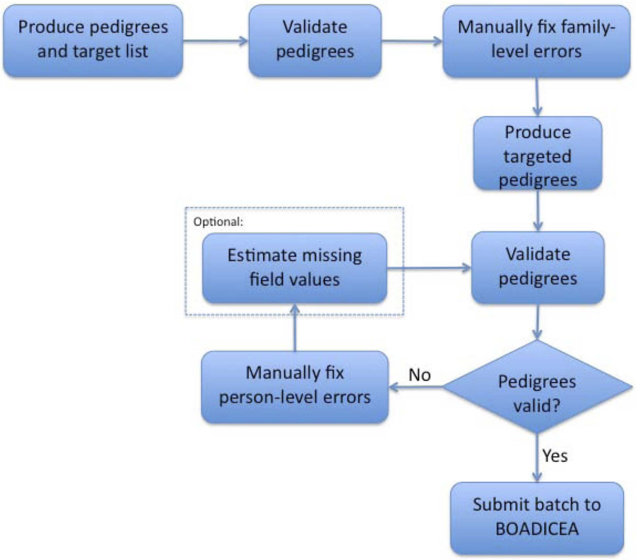
A.3 Validating pedigrees. BOADICEACentre checks the validity of the data in each file and generates a report that describes the types of errors detected and lists the affected individuals. At present, it can detect up to 45 different types of syntactic (e.g., wrong field delimiter) and semantic (e.g., a pedigree member older than his or her living parent) pedigree errors. A file is produced containing the filenames of all pedigrees that did not pass validation. This list can be input into other BOADICEACentre components, most notably, the risk estimation component (see A.5).

A.4 Missing field imputation. BOADICEACentre uses a rule-based algorithm to impute missing year of birth and age at last follow-up. These fields are mandatory for any individual who is either the target for risk estimation or affected by any type of cancer. The algorithm estimates missing values based on the mean values of a person's relatives. Refer to Dite *et al* (2003, 2010) for more detailed descriptions of the imputation rules.

Imputation is achieved using an iterative estimation process; estimates of later iterations may be based on the estimates calculated in previous iterations.

A.5 Obtaining risk estimates for a batch. BOADICEACentre submits the pedigree files to BOADICEA via the internet and collates the results into a single spreadsheet of results.

BOADICEACentre's requires that the user has an existing BOADICEA web client account and that the pedigree files have passed validation (see A.3). Pedigree files that encounter a BOADICEA run-time error will not be processed. The user may optionally provide a list of targeted pedigrees to exclude from submission (e.g., a list of pedigrees that failed validation). Refer to the below figure for a typical BOADICEACentre workflow.


Produce pedigrees and target list
Validate pedigrees
Manually fix family-level errors
Produce targeted pedigrees
Optional:
Estimate missing field values
Validate pedigrees
Manually fix person-level errors
No
Pedigrees valid?
Yes
Submit batch to BOADICEA