Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study

Ben Van Calster[1], *professor*, Kirsten Van Hoorde[2,3], *doctoral researcher*, Lil Valentin[4], *professor*, Antonia C. Testa[5], *professor*, Daniela Fischerova[6], *consultant gynaecologist*, Caroline Van Holsbeke[7], *consultant gynaecologist*, Luca Savelli[8], *consultant gynaecologist*, Dorella Franchi[9], *consultant gynaecologist*, Elisabeth Epstein[10], *professor*, Jeroen Kaijser[1,11], *research fellow*, Vanya Van Belle[2,3], *postdoctoral researcher*, Artur Czekierdowski[12], *consultant gynaecologist*, Stefano Guerriero[13], *professor*, Robert Fruscio[14], *consultant gynaecologist*, Chiara Lanzani[15], *consultant gynaecologist*, Felice Scala[16], *consultant gynaecologist*, Tom Bourne[1,11,17], *professor*, Dirk Timmerman[1,11], *professor*; International Ovarian Tumour Analysis (IOTA) group

[1] KU Leuven Department of Development and Regeneration, Herestraat 49 box 7003, 3000 Leuven, Belgium; [2] KU Leuven Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium; [3] KU Leuven iMinds Medical Information Technologies, Kasteelpark Arenberg 10, 3001 Leuven, Belgium; [4] Department of Obstetrics and Gynaecology, Skåne University Hospital Malmö, Lund University, Södra Förstadsgatan, 20502 Malmö, Sweden; [5] Department of Oncology, Catholic University of the Sacred Heart, Largo Francesco Vito 8, Rome, 00165, Italy; [6] Gynaecological Oncology Center, Department of Obstetrics and Gynaecology, Charles University, Apolinarska 18, 120 00 Prague, Czech Republic; [7] Department of Obstetrics and Gynaecology, Ziekenhuis Oost Limburg, Schiepse Bos 6, 3600 Genk, Belgium; [8] Gynaecology and Reproductive Medicine Unit, S. Orsola-Malpighi Hospital, University of Bologna, Via Albertoni 15, Bologna 40138, Italy; [9] Preventive Gynaecology Unit, Division of Gynaecology, European Institute of Oncology, via Ripamonti 435, Milan 20141, Italy; [10] Department of Obstetrics and Gynaecology, Karolinska University Hospital, Karolinskavägen (Solna), SE-17176 Stockholm, Sweden; [11] Department of Obstetrics and Gynaecology, University Hospitals Leuven, Herestraat 49 box 7003, 3000 Leuven, Belgium; [12] Ist Department of Gynaecological Oncology and Gynaecology, Medical University in Lublin, ul. Staszica 16, 20-081 Lublin, Poland; [13] Department of Obstetrics and Gynaecology, Azienda Ospedaliero Universitaria di Cagliari, Strada Statale 554 Monserrato, Cagliari 09045, Italy; [14] Clinic of Obstetrics and Gynaecology, University of Milan-Bicocca, San Gerardo Hospital, Via Pergolesi, 33, Monza 20052, Italy; [15] Department of Woman, Mother and Neonate, Buzzi Children's Hospital, Biological and Clinical School of Medicine, University of Milan, Via Castelvetro 32, 20154, Milan, Italy; [16] Department of Gynaecologic Oncology, Istituto Nazionale Tumori, Via M. Semmola, Naples 80131, Italy; [17] Queen Charlotte's and Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, United Kingdom

WORD COUNT 3716 (excluding Appendix)

Corresponding author:
Ben Van Calster, PhD
Department of Development and Regeneration
KU Leuven
Herestraat 49 box 7003
B-3000 Leuven
Belgium
ben.vancalster@med.kuleuven.be

Appendix A: Multiple imputation of missing values for serum CA-125

In the available data, the serum CA-125 level was missing in 31% of the women. The value was more often missing in women with a benign tumour (36%) than in women with a borderline (18%), stage I invasive (20%), stage II-IV invasive (17%), or secondary metastatic (25%) tumour. The missing values are highly likely to have occurred for two reasons. First, following different local management practices some centres were more committed than others to measure CA-125. Second, investigators sometimes refrained from measuring CA-125 based on the overall clinical picture and the appearance of the tumour on ultrasound. The percentage of missing values per centre varied between 0% and 94% with a median of 19%. Seventeen of 24 centres had at most 30% missing values, and four centres had clearly more than 50% missing values: Bologna (N=348, 69% missing values), Genk (N=428, 72% missing values), Barcelona (N=37, 84% missing values), Maurepas (N=64, 94% missing values).

We used multiple imputation to handle the missing values in the analysis[1]: the missing CA-125 values were estimated (i.e. imputed) multiple times to acknowledge the uncertainty in the imputed values. To estimate the missing values, we used predictive mean matching regression[2] using variables that were related to either the level of CA-125 itself, or to the unavailability of CA-125 (i.e. a binary indicator indicating for each woman whether CA-125 was missing or not). This was repeated 100 times to generate multiple imputations of the missing values, resulting in 100 completed data sets. To reduce computational burden, we used only five of the completed datasets for variable selection. Differences in results between these five datasets were minimal. Subsequent model fitting was carried out on all 100 completed datasets and the resulting coefficients were averaged to obtain a final set of model coefficients.

We consider multiple imputation more appropriate than an analysis including only patients with information on CA-125 (complete case analysis). A complete case analysis will introduce bias: (1) because CA-125 is more often missing in patients with a tumour judged to be benign based on its ultrasound appearance, estimated risks for malignant tumour types will be too high, and (2) exclusion of these "easy" tumours will make the complete case sample more homogeneous and will therefore decrease model performance. In addition to introducing bias, a complete case analysis would waste large amounts of precious data because there is complete information on all variables except CA-125.

Appendix B: Variable selection

Variable selection was performed in two stages. To avoid overfitting, we relied on subject matter knowledge to reduce the number of potential predictors.[3,4] Clinicians with experience in diagnosing adnexal masses judged the available variables with respect to 'measurement harm' as well as their likely ability to differentiate between the four subtypes of malignant tumours. Measurement harm refers to issues such as subjectivity, dependency on operator experience, and burden for the patient. In addition we estimated to what extent the values of the variables varied between centres after controlling for tumour histology, in order to favour predictors with limited heterogeneity.[5] Finally, we aimed to include some baseline information on centre type, in order to improve the calibration of the estimated probabilities. All of these issues were taken into account when selecting variables. We selected four clinical variables, i.e. age (years), serum CA-125 (U/mL), family history of ovarian cancer (yes/no), and type of centre (oncology centre vs other hospitals), and six ultrasound variables, i.e. the maximum diameter of the lesion (mm), proportion of solid tissue (i.e. maximum diameter of the largest solid component divided by the maximum diameter of the lesion), the presence of more than 10 cyst locules (yes/no), number of papillary projections (0, 1, 2, 3, more than 3), presence of acoustic shadows (yes/no), and presence of ascites (yes/no). We believe that this a priori selection was sensible because it was done through collaboration between four experienced clinical researchers (DT, TB, CVH, LV).

The second stage of the variable selection procedure involved limited data-driven variable selection based on the method of multivariable fractional polynomials, which simultaneously selects variables and determines the optimal transformation of numerical variables using fractional polynomials.[6] The variable selection procedure (called RA2) is a variant of the standard backward selection procedure. The RA2 procedure was

performed separately for the four submodels of the multinomial logistic regression model: Borderline versus Benign, Stage I cancer versus Benign, Stage II-IV cancer versus Benign, and Secondary metastatic cancer versus Benign. The variable selection analysis was carried out using logistic regression analysis with random centre intercepts. As elimination criterion in the backward selection process and as criterion for the selection of functional form we used the p-value that mimicked the Bayesian Information Criterion,[4] in order to end up with a parsimonious model. We allowed a first degree polynomial for age, serum CA-125, the maximum diameter of the lesion, and the number of papillary projections. For the proportion of solid tissue a second degree polynomial was allowed because this is considered an important predictor that is of semi-continuous nature because a large number of tumours do not have a solid component. The final set of predictors for the polytomous model consisted of the variables that were selected in any of the four submodels. If different transformations were selected in the submodels, one final transformation was chosen. Two predictors were forced into the model and hence could not be eliminated: age and type of centre.

The variable “type of centre” was included because the risk of malignancy is likely to be higher in oncology centres than in other centres even after adjusting for clinical and ultrasound characteristics. We chose to add type of centre as a predictor rather than to develop separate models for oncology centres and other hospitals. Developing separate models means that sample size is reduced, mainly for the non-oncology hospitals. This would increase the possibility of overfitting. Developing one model with type of centre as a predictor assumes that the effects of clinical and ultrasound predictors are the same in both types of centres, an assumption that we consider plausible.

Appendix C: Shrinkage factors

We multiplied the predictor coefficients with uniform 'shrinkage factors' to avoid exaggerated model coefficients.[4,7] These factors were estimated as follows. For every submodel (Borderline versus Benign, Stage I cancer versus Benign, Stage II-IV cancer versus Benign, and Secondary metastatic cancer versus Benign) a separate random intercepts logistic regression model was fitted using the selected predictors. If M is the improvement in the minus 2 log-likelihood over a random intercepts only model, the shrinkage factor is obtained as $(M - df)/M$, where the df are the degrees of freedom spent for the candidate predictors. In our situation df is 17: one for each of the five binary predictors (family history of ovarian cancer, type of centre, >10 cyst locules, acoustic shadows, and ascites), two for the four predictors for which a first degree polynomial was allowed in the multivariable fractional polynomial procedure (age, serum CA-125, maximum diameter of the lesion, and number of papillary projections), and four for the predictor for which a second degree polynomial was allowed (proportion of solid tissue).[6]

After fitting the random intercepts multinomial logistic regression model, we multiplied the obtained coefficients with the appropriate shrinkage factor. Then, we re-estimated model intercepts accordingly such that the average value of the linear predictors after shrinkage equal the average value of the fitted model without shrinkage.

Appendix D: The formula of the ADNEX model after retraining on the pooled data

The ADNEX model uses the following predictors, with measurement unit between parentheses and reference letter between square brackets: patient age (years) [A], serum CA-125 (U/mL) [B], maximal diameter of lesion (mm) [C], maximal diameter of largest solid component (mm) [D], more than 10 cyst locules (1-yes vs 0-no) [E], number of papillary structures (0, 1, 2, 3, 4; with 4 indicating more than three) [F], acoustic shadows (1-yes vs 0-no) [G], ascites (1-yes vs 0-no) [H], and type of centre (1-oncology centre vs 0-other centre types) [I].

The linear predictors z_1 to z_4 contain model coefficients for each predictor for the prediction of borderline vs. benign tumours (z_1), stage I cancer vs. benign tumours (z_2), stage II-IV cancer vs. benign tumours (z_3), and secondary metastatic cancer vs. benign tumours (z_4). The linear predictors are as follows:

$$z_1 = -7.577663 + 0.004506 * A + 0.111642 * \text{Log2}(B) + 0.372046 * \text{Log2}(C) + 6.967853 * (D/C) - 5.65588 * (D/C)^2 + 1.375079 * E + 0.604238 * F - 2.04157 * G + 0.971061 * H + 0.953043 * I$$

$$z_2 = -12.276041 + 0.017260 * A + 0.197249 * \text{Log2}(B) + 0.873530 * \text{Log2}(C) + 9.583053 * (D/C) - 5.83319 * (D/C)^2 + 0.791873 * E + 0.400369 * F - 1.87763 * G + 0.452731 * H + 0.452484 * I$$

$$z_3 = -14.915830 + 0.051239 * A + 0.765456 * \text{Log2}(B) + 0.430477 * \text{Log2}(C) + 10.37696 * (D/C) - 5.70975 * (D/C)^2 + 0.273692 * E + 0.389874 * F - 2.35516 * G + 1.348408 * H + 0.459021 * I$$

$$z_4 = -11.909267 + 0.033601 * A + 0.276166 * \text{Log2}(B) + 0.449025 * \text{Log2}(C) + 6.644939 * (D/C) - 2.30330 * (D/C)^2 + 0.899980 * E + 0.215645 * F - 2.49845 * G + 1.636407 * H + 0.808887 * I$$

The probabilities of the five types of tumour are computed as follows:

$$P(\text{Benign}) = \frac{1}{1+\exp(z_1)+\exp(z_2)+\exp(z_3)+\exp(z_4)}$$

$$P(\text{Borderline}) = \frac{\exp(z_1)}{1+\exp(z_1)+\exp(z_2)+\exp(z_3)+\exp(z_4)}$$

$$P(\text{Stage I cancer}) = \frac{\exp(z_2)}{1+\exp(z_1)+\exp(z_2)+\exp(z_3)+\exp(z_4)}$$

$$P(\text{Stage II-IV cancer}) = \frac{\exp(z_3)}{1+\exp(z_1)+\exp(z_2)+\exp(z_3)+\exp(z_4)}$$

$$P(\text{Metastatic cancer}) = \frac{\exp(z_4)}{1+\exp(z_1)+\exp(z_2)+\exp(z_3)+\exp(z_4)}$$

Appendix reference list

1. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.

2. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. Comput Stat Data Anal 1996;22:425-46.

3. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001.

4. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer, 2009.

5. Wynants L, Timmerman D, Bourne T, Van Huffel S, Van Calster B. Screening for data clustering in multicenter studies: the residual intraclass correlation. BMC Med Res Methodol 2013;13:128.

6. Royston P, Sauerbrei W. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, 2008.

7. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303-25.

Table S1. Prevalence of specific pathologies in the pooled data (N=5909) and separately for patients from oncology centres and other hospitals.

Tumour pathology	**Full dataset, N (%)**	**Oncology centres, N (%)**	**Other hospitals, N (%)**
Benign			
Endometrioma	1055 (17.9)	551 (14.7)	504 (23.3)
Benign teratoma (dermoid)	633 (10.7)	369 (9.8)	264 (12.2)
Simple/parasalpingeal cyst	386 (6.5)	197 (5.3)	189 (8.8)
Functional cyst	156 (2.6)	81 (2.2)	75 (3.5)
Hydrosalpinx	147 (2.5)	68 (1.8)	79 (3.7)
Peritoneal pseudocyst	39 (0.7)	23 (0.6)	16 (0.7)
Abscess	59 (1.0)	40 (1.1)	19 (0.9)
Fibroma	282 (4.8)	186 (5.0)	96 (4.4)
Serous cystadenoma	679 (11.5)	370 (9.9)	309 (14.3)
Mucinous cystadenoma	453 (7.7)	236 (6.3)	217 (10.0)
Rare benign pathologies	91 (1.5)	57 (1.5)	34 (1.6)
Malignant			
Primary invasive stage I	263 (4.5)	211 (5.6)	52 (2.4)
Primary invasive stage II	94 (1.6)	72 (1.9)	22 (1.0)
Primary invasive stage III	731 (12.4)	626 (16.7)	105 (4.9)
Primary invasive stage IV	119 (2.0)	99 (2.6)	20 (0.9)
Rare primary invasive pathologies	137 (2.3)	92 (2.5)	45 (2.1)
Bordeline stage I	299 (5.1)	228 (6.1)	71 (3.3)
Bordeline stage II	14 (0.2)	9 (0.2)	5 (0.2)
Bordeline stage III	25 (0.4)	23 (0.6)	2 (0.1)
Bordeline stage IV	1 (<0.1)	1 (<0.01)	0 (0)
Secondary metastatic cancer	246 (4.2)	210 (5.6)	36 (1.7)

Table S2. Demographic and reproductive characteristics by tumour type for the pooled dataset (N=5909).

Variable	Benign	Borderline	Stage I	Stage II-IV	Metastatic
	N=3980	N=339	N=356	N=988	N=246
Age (years)	42	49	54	59	57
Postmenopausal	30.1%	42.8%	57.6%	72.5%	65.9%
Parity[a]	1	1	1	2	2
Nulliparous[a]	44.5%	34.9%	30.7%	20.3%	23.3%
Personal history of ovarian cancer	0.9%	11.2%	1.1%	1.6%	2.4%
Family history of ovarian cancer,	2.0%	3.0%	3.7%	5.8%	2.0%
Previous hysterectomy	6.3%	5.9%	9.8%	6.5%	6.9%
Current use of hormonal therapy	14.9%	10.6%	7.3%	7.1%	10.6%
Personal history of breast cancer[a]	3.2%	2.2%	7.4%	7.9%	15.8%
Family history of breast cancer[a]	8.6%	12.9%	14.2%	13.1%	9.2%

Statistics shown are median for age and parity, and percentage for categorical variables.

[a] Results based on the development data (N=2557, N=186, N=176, N=467, N=120 for the five tumour types) because this information was not collected in phase 3 of the IOTA study (validation data).

Table S3. Discrimination results on the validation data (n=2403) and after retraining of the model on the pooled data (n=5909). The results are shown separately for oncology centres and other hospitals.

	Oncology centres		Other hospitals	
Performance measures	**Validation**	**After retraining on pooled data**	**Validation**	**After retraining on pooled data**
AUC Benign vs Malignant	0.94	0.94	0.93	0.95
AUC Benign vs Borderline	0.85	0.87	0.84	0.88
AUC Benign vs Stage I	0.90	0.92	0.93	0.94
AUC Benign vs Stage II-IV	0.98	0.98	0.99	0.99
AUC Benign vs Metastatic	0.94	0.95	0.97	0.97
AUC Borderline vs Stage I	0.74	0.74	0.83	0.82
AUC Borderline vs Stage II-IV	0.94	0.93	0.97	0.92
AUC Borderline vs Metastatic	0.87	0.88	0.94	0.89
AUC Stage I vs Stage II-IV	0.86	0.85	0.87	0.83
AUC Stage I vs Metastatic	0.70	0.74	0.75	0.76
AUC Stage II-IV vs Metastatic	0.82	0.81	0.81	0.71
Polytomous Discrimination Index	0.56	0.56	0.56	0.56

AUC, area under the receiver operating characteristic curve.
For five tumour types, the Polytomous Discrimination Index for random prediction equals 0.2, hence its value cannot be directly compared with the AUCs.

Table S4. Discrimination performance of the model without CA-125 as a predictor as obtained on the validation data and after retraining on the pooled data.

Performance measures	**Validation data (n=2403)**	**After retraining on pooled data (n=5909)**
AUC Benign vs malignant	0.932 (0.922 to 0.941)	0.940 (0.934 to 0.946)
AUC Benign vs Borderline	0.85 (0.81 to 0.87)	0.88 (0.86 to 0.90)
AUC Benign vs Stage I	0.91 (0.89 to 0.93)	0.93 (0.92 to 0.94)
AUC Benign vs Stage II-IV	0.97 (0.96 to 0.97)	0.97 (0.96 to 0.97)
AUC Benign vs Metastatic	0.95 (0.93 to 0.96)	0.96 (0.94 to 0.97)
AUC Borderline vs Stage I	0.76 (0.69 to 0.80)	0.76 (0.72 to 0.79)
AUC Borderline vs Stage II-IV	0.91 (0.88 to 0.93)	0.90 (0.88 to 0.92)
AUC Borderline vs Metastatic	0.87 (0.82 to 0.91)	0.89 (0.86 to 0.91)
AUC Stage I vs Stage II-IV	0.76 (0.72 to 0.80)	0.75 (0.72 to 0.78)
AUC Stage I vs Metastatic	0.70 (0.64 to 0.76)	0.74 (0.70 to 0.78)
AUC Stage II-IV vs Metastatic	0.59 (0.53 to 0.64)	0.62 (0.59 to 0.66)
Polytomous Discrimination Index	0.494 (0.467 to 0.509)	0.511 (0.494 to 0.526)

AUC, area under the receiver operating characteristic curve.
For five tumour types, the Polytomous Discrimination Index for random prediction equals 0.2, hence its value cannot be directly compared with the AUCs.
95% confidence intervals are shown in parentheses

Figure S1. Forest plot with centre-specific areas under the receiver operating characteristic curve (AUC) with regard to discrimination between benign and malignant tumours. Results for the validation data are presented. The AUC is consistent over centres, notwithstanding the variability observed for centres that contributed few patients. NC, not computed.

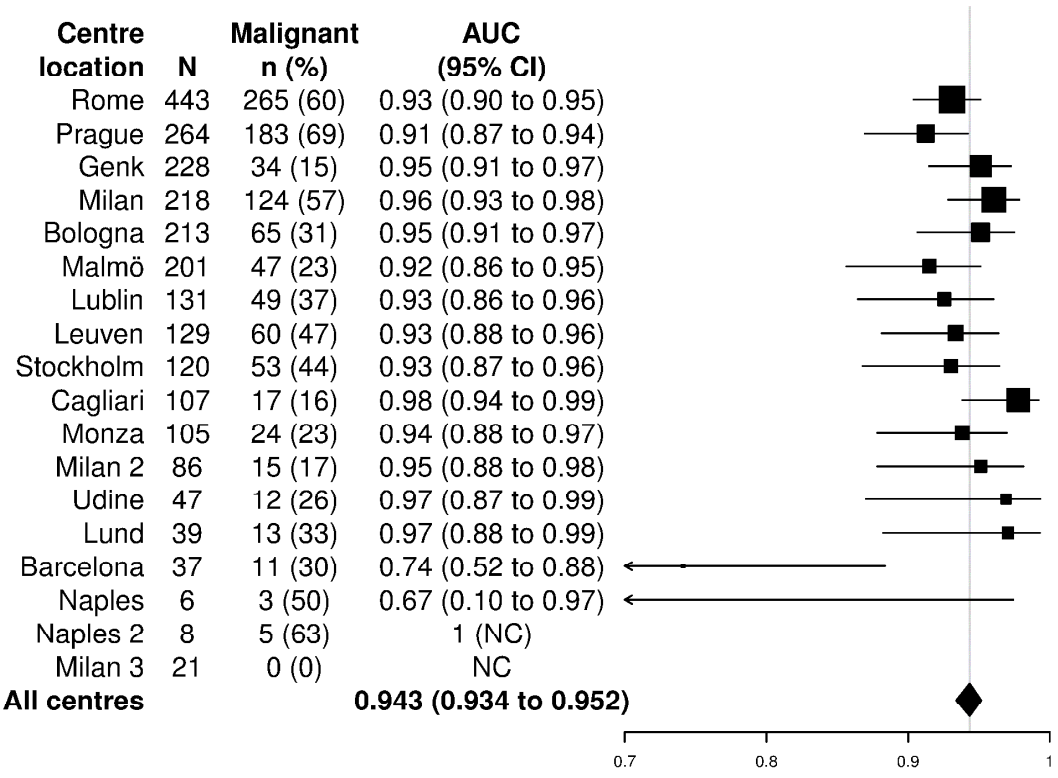

Figure S2. Calibration plots of the predicted risks for each type of tumour in oncology centres. Data have been calculated using the validation data (n=1715). The plots show how well the predicted probabilities (x-axis) agree with the observed probabilities (y-axis). For perfect agreement, the calibration curve falls on the ideal diagonal line. The histograms below the plots show the distribution of the estimated risks.

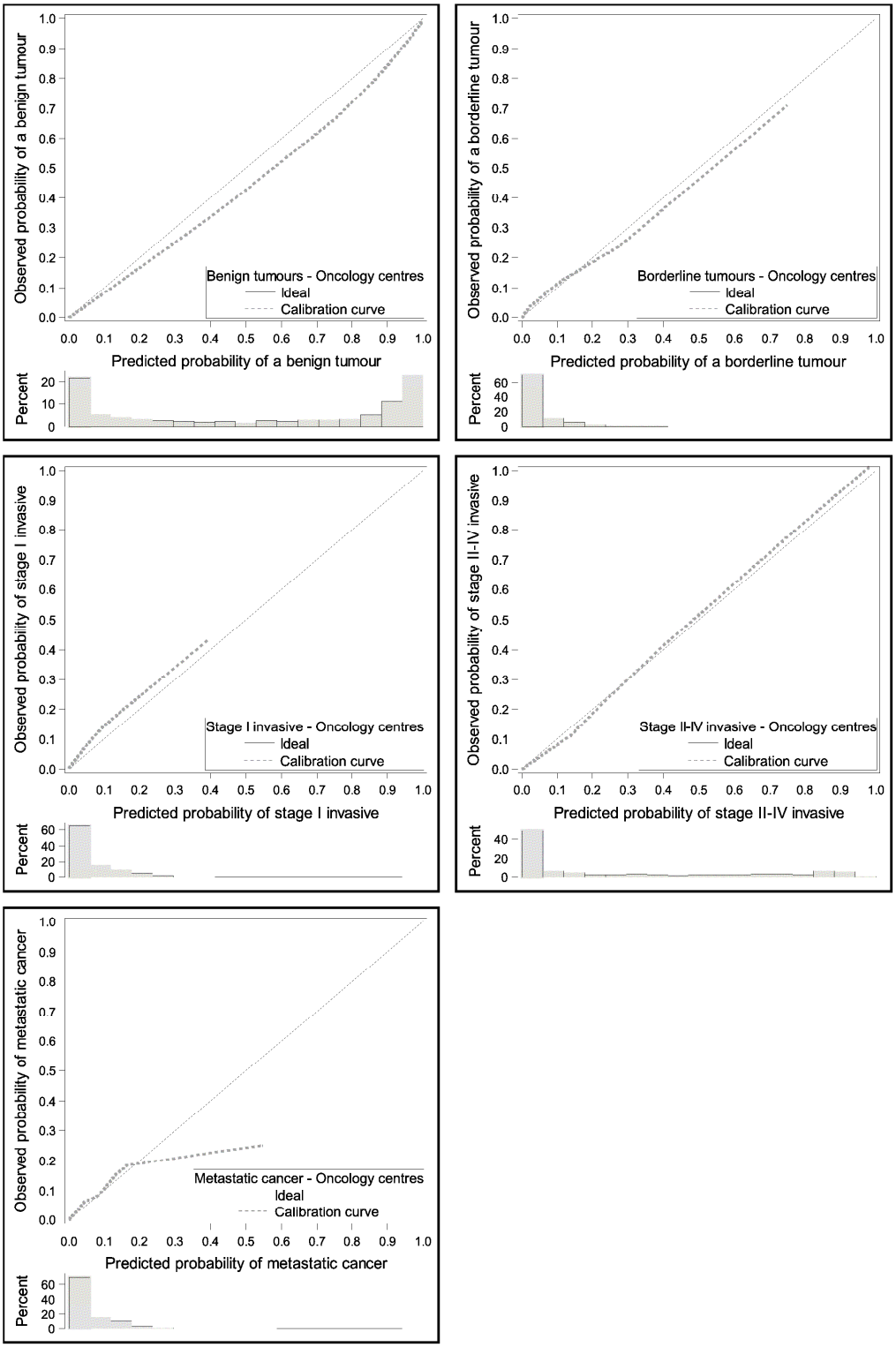

Figure S3. Calibration plots of the predicted risks for each type of tumour in other than oncology centres. Data have been calculated using the validation data (n=688). The plots show how well the predicted probabilities (x-axis) agree with the observed probabilities (y-axis). For perfect agreement, the calibration curve falls on the ideal diagonal line. The histograms below the plots show the distribution of the estimated risks. For this subgroup the results contain a fair degree of uncertainty due to limited number of malignancies.