Simple dichotomous updating methods improved the validity of polytomous models

K. Van Hoorde[1,2], Y. Vergouwe[3], D. Timmerman[4],

S. Van Huffel[1,2], E.W. Steyerberg[3], B. Van Calster[4]

1 Department of Electrical Engineering (ESAT-SCD), KU Leuven – University of Leuven, Leuven, Belgium

2 iMinds Future Health Department, KU Leuven – University of Leuven, Leuven, Belgium

3 Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

4 Department of Development & Regeneration, KU Leuven – University of Leuven, Leuven, Belgium

Corresponding author:

Kirsten Van Hoorde

Department of Electrical Engineering (ESAT-SCD)

KU Leuven – University of Leuven

Kasteelpark Arenberg 10 - box 2446

3001 Leuven

Belgium

Tel.: +32 16 321160

Fax: +32 16 321970

E-mail address: kirsten.vanhoorde@esat.kuleuven.be

Abstract (199 words)

Objective: Prediction models may perform poorly in a new setting. We aimed to determine which model updating methods should be applied for models predicting polytomous outcomes, which often suffer from one or more categories with low prevalence.

Study Design and Setting: We used case studies on testicular and ovarian tumors. The original regression models were based on 502 and 2037 patients, and validated on 273 and 1107 patients, respectively. The polytomous models combined dichotomous models for category A versus B+C and B versus C ('sequential dichotomous modeling'). Simple recalibration, revision, and redevelopment methods were considered. To assess discrimination (using dichotomous and polytomous c-statistics) and calibration (by comparing observed and expected prevalences) of these methods, the validation data were divided into updating and test parts. Five hundred such divisions were randomly generated, and the average test set results reported.

Results: None of the updating methods could improve discrimination of the original models, but recalibration, revision, and redevelopment strongly improved calibration. Redevelopment was unstable with respect to overfitting and performance.

Conclusion: Simple dichotomous updating methods behaved well when applied to polytomous models. Our results suggest that recalibration is preferred, but larger validation sets may make revision or redevelopment a sensible alternative.

Keywords: prediction models, model updating, polytomous outcomes, sequential dichotomous modeling, discrimination, calibration

Running title: Updating polytomous risk prediction models

Word count: 3454

What is new?

Key findings: Updating methods for prediction models with dichotomous outcomes show similar behavior when applied to prediction models with polytomous outcomes. Simple recalibration methods were preferred over no updating or redevelopment. Sample size issues are common for polytomous outcomes, such that revision methods can often not be reliably implemented even when predictor effects are different in the updating population.

What the study adds to what was known: Existing updating methods can be used for polytomous prediction models. Similar to the updating of dichotomous models, simple methods are recommended, especially when the amount of data available for updating is limited. When sample size is large, more complex methods such as revision or even redevelopment can be a valid alternative.

Implications of the study: To obtain more accurate predictions, polytomous prediction models should be updated when performance in a new setting is poor. The preferred updating method depends on the amount of data available. Because of these findings, further research will focus on updating methods for multinomial logistic regression, which is the most common approach for polytomous outcomes.

1. Introduction

Prediction models are important for personalized health care and to assist with clinical decision making. Therefore these models have to be sufficiently accurate. This means that a model discriminates well between patients with and without the disease of interest. Another important aspect is calibration of the risk estimates: the predicted risks should correspond with observed proportions of the endpoint [1].

Validation of developed prediction models is essential, especially external validation in time or place [2-4]. The external validity of a prediction model can be disappointing, with decreased performance when tested in a new center or a different setting. This often leads researchers to discard the model and to develop a new one. However, model updating is often a sensible alternative [5,6,7]. Despite a model's disappointing performance, it may still contain useful information with respect to relevant predictors and predictor effects. It is recognized that it may be more efficient to update an existing prediction model than to start the development anew, especially in relatively small validation sets [5]. This approach is similar to the use of prior information in Bayesian analysis [8]. Updating an existing model means that we combine information from the original model with information in the new data.

Sample size issues are important when the outcome to be predicted is polytomous, because such outcomes often suffer from categories with low prevalence. The aim of the present study is to assess whether and how previously suggested updating techniques for dichotomous models translate to the setting of polytomous outcome prediction [5,9,10]. The methods are demonstrated on two case studies dealing with testicular cancer and ovarian tumors. To apply the existing updating methods, we use sequential dichotomous modeling

[11,12] to obtain polytomous prediction models. If results are positive, further research will extend updating methods to prediction models based on multinomial logistic regression, as this is the most common approach for polytomous outcomes.

2. Patient and Methods

2.1. Study set-up

We applied the updating methods on two case studies, one on predicting residual mass histology after chemotherapy for testicular cancer, and one on predicting ovarian tumor pathology. We derived a prediction model on the development data, and applied it to validation data from a different setting. The model was updated with a number of updating methods. To investigate the performance of the updating methods, the validation data were divided at random into an updating set and a test set on which the updated model was evaluated. This division was repeated 500 times and the average over test sets calculated. SAS version 9.2 and R version 2.15 were used for the statistical analysis.

2.2. Case studies

Testicular cancer. After patients have been treated with cisplatin-based chemotherapy for advanced nonseminomatous testicular germ cell cancer, retroperitoneal lymph nodes may still contain tumor cells [13,14]. An accurate prediction of the residual mass containing necrosis or fibrosis (benign tissue), teratoma, or viable cancer is important to decide whether or not the mass should be surgically resected (teratoma or viable cancer), and whether additional chemotherapy is indicated (viable cancer). The development dataset consisted of resected masses of patients recruited from different hospitals in three countries

[14]. Among 502 patients, 237 (47%) had benign masses, 213 (42%) a teratoma and 52 (10%) viable cancer (Table 1). The prediction model was updated on data from a tertiary referral center from Indiana University. The validation data consist of 273 resected masses (Table 1), of which 76 (28%) were benign, 162 (59%) were teratomas and 35 (13%) were viable cancers [14].

Ovarian tumor. When an ovarian tumor is detected on ultrasound examination, an accurate diagnosis of the tumor as benign, borderline malignant, or invasive is crucial for selecting the optimal management strategy for each patient [15]. In case of invasive cancer, the type and quality of the surgery affects prognosis. For borderline malignancies, less invasive surgical methods are typically preferred, for example fertility-sparing methods in young women. In most benign asymptomatic tumors conservative management is a valid option. If surgery is indicated because of symptoms or at patient request minimally invasive surgery is the preferred management in most patients with benign tumors. The data contained patients with an ovarian tumor that was considered sufficiently suspicious of possible malignancy to warrant surgical intervention according to local management protocols. Development data were collected at tertiary referral centers for oncology in seven countries within the framework of the International Ovarian Tumor Analysis (IOTA) study [16-18]. The dataset contained 2037 women including 1315 (65%) with a benign tumor, 139 (7%) with a borderline malignancy and 583 (29%) with invasive cancer (Table 1). The prediction model was updated on data from regional or public hospitals in five countries, also as part of the IOTA study. The validation data contained 1107 patients (918 (83%) with a benign tumor, 41 (4%) with a borderline malignancy and 148 (13%) with invasive cancer) (Table 1).

2.3. Polytomous prediction models

We obtained polytomous prediction models using a sequential dichotomous modeling strategy (also labeled consecutive dichotomous modeling, binary tree modeling, or nested dichotomies) [11,12]. Roukema and colleagues observed similar performance for polytomous logistic regression and sequential dichotomous logistic regression [12]. For outcomes with three categories as in our examples, this involves the development of two models: one for category A versus B+C and one for category B vs C. The risk estimates for categories B or C were obtained by multiplying the risk for B+C from the first model with the risk for category B or category C from the second model. Logistic regression was used to develop the dichotomous models. In general for k categories there will be k-1 dichotomous models. There are different ways to combine categories (e.g. A vs B-K or A-B vs C-K). Where possible, the choice of which categories to combine at each stage should depend on clinical judgment based on differences between categories concerning treatment consequences and clinical profile. Category prevalence may also play a role.

Testicular cancer. The first model discriminated between benign and malignant (teratoma or viable cancer) resected masses, the second model between teratomas and viable cancers [19]. The models estimated the probability of a benign mass and of a teratoma, respectively. For both models five predictor variables were considered: presence of teratoma elements in the primary tumor (yes/no), prechemotherapy levels of the serum tumor markers α-fetoprotein (AFP) and human chorionic gonadotrophin (hCG) (normal/elevated), the square root of the maximum diameter of the residual mass measured on computerized tomography (CT) after chemotherapy, and the change in diameter induced by chemotherapy (Table 1). Backward variable selection with Akaike's Information Criterion (AIC) as stopping rule was used to develop the prediction models. The first model (benign vs malignant) included all

five predictors whereas the second model (teratoma vs viable cancer) contained four predictors after dropping the prechemotherapy level of the serum tumor marker AFP (see Table S1 for model coefficients).

Ovarian tumor. The first model discriminated between benign and malignant (borderline or invasive) tumors, the second model between borderline and invasive tumors. The models estimated the probability of a benign tumor and of a borderline tumor, respectively. Seven variables were considered: age, square root of the ratio of the maximum diameters of the largest solid component and the lesion, presence of masses on both ovaries (yes/no), presence of ascites (yes/no), irregular internal cyst wall (yes/no), current use of hormonal therapy (yes/no) and presence of acoustic shadows (yes/no) (Table 1). After backward variable selection with AIC as stopping rule, the dichotomous logistic regression model for benign versus malignant contained all seven variables. The second dichotomous model contained five variables and excluded current use of hormonal therapy and presence of acoustic shadows (Table S1).

2.4. Updating methods

We considered six different updating methods (Table 2) [5], which were applied to the two sequential dichotomous models on the updating dataset. Thus, none of these methods used the original development dataset to update the prediction model. The methods varied in the number of parameters that were (re-)estimated.

The simplest method (method 0) uses the original model without any adjustment, and therefore represents a reference method with which the other methods are compared. Methods 1 and 2 are recalibration methods. Method 1 only updates the model's intercept to

correct the calibration-in-the-large, i.e. to take care that the mean predicted probability equals the observed prevalences. Method 2 updates the intercept and uses the calibration slope as a single overall adjustment factor to update the regression coefficients. Methods 3 and 4 are more extensive revision methods. Method 3 first uses the approach from method 2 to update the model coefficients. Then likelihood ratio tests are used for each predictor to assess whether the effect in the updating dataset is statistically significantly different from the effect obtained using method 2. If so, the effect of the predictor is re-estimated. A forward re-estimation procedure is followed: we first consider the predictor with the strongest difference, and continue until all tests fail to reach statistical significance. For method 4 we re-estimate the intercept and individual model regression coefficients for all predictors on the updating dataset. For methods 3 and 4 parameter shrinkage is used to prevent overfitting. We used a simple heuristic shrinkage factor defined as (model χ^2 - df) / model χ^2 [5,20]. Model χ^2 refers to the difference in -2 log likelihood between the model with re-estimated predictors and the re-calibrated model (method 2), and df corresponds to the difference in degrees of freedom of these two models. By multiplying all regression coefficients with the shrinkage factor, they are pulled towards their re-calibrated values [5,9,20]. The intercept of the shrunken model is re-estimated to ensure that the sum of predicted probabilities equals the sum of observed outcomes [5]. Finally, method 5 builds a completely new model with exactly the same procedure as used for developing the original model (i.e. including backward variable selection).

2.5 Performance evaluation, events per variable and simulations

We assessed the discrimination and calibration performance of the polytomous risk predictions for each updating method. Discrimination was assessed with the polytomous

discrimination index (PDI) [21]. Discrimination was assessed with the polytomous discrimination index (PDI). Considering a set of patients with one patient per outcome category, the PDI counts for how many outcome categories the predicted probability is highest for the patient belonging to that category. PDI is obtained as the average count over all possible groups of cases, divided by the number of outcome categories. PDI hence estimates the probability to correctly identify the patient from a randomly selected outcome category from a group of patients who all belong to a different category [21]. PDI corresponds to 1/k in case of random performance [21]. Discrimination for the two sequential dichotomous models was assessed with the standard c-statistic, which is equivalent to the area under the ROC curve. The c-statistic in case of random performance corresponds to 0.5. The calibration is assessed using calibration-in-the-large: we determined for each outcome category the difference between the observed prevalence and the average predicted risk (i.e. the expected prevalence).

Given that one updating method involves the development of a new model on the updating dataset, we decided to vary the number of events per variable (EPV) in the updating set to assess its influence. We calculated EPV by dividing the size of the smallest outcome category of the updating set by the number of considered variables in the model development process, thereby acknowledging that this ignores the fact that each variable is considered in two dichotomous models. To vary EPV we considered different sample sizes for the updating dataset, while the size of the test set was kept the same. We focused on low EPV situations (EPV 5 or 3). For the testicular cancer case study we considered five predictors, such that the smallest category in the updating set has 25 events for EPV 5 (5 predictors * 5 events per predictor = 25) and 15 for EPV 3 (5*3=15). For the ovarian tumor study we considered seven

predictors, such that the smallest category in the updating set has 35 events for EPV 5 (7*5=35) and 21 for EPV 3 (7*3=21). The random division in updating and test data was stratified by outcome category (Table S2). The division of the validation set into updating and test set was repeated 500 times, and average test set results were reported.

3. Results

Performance of original models in development data

For the testicular cancer study the c-statistic for modeling benign vs malignant masses was 0.81 [95% CI: 0.77-0.85], and for modeling teratoma vs viable cancer 0.65 [95% CI: 0.56-0.74]. The PDI of the polytomous model was 0.57 [95% CI: 0.52-0.62]. For the ovarian tumor study the c-statistic was 0.91 [95% CI: 0.90-0.93] for modeling benign vs malignant tumors and 0.84 [95% CI: 0.81-0.88] for modeling borderline vs invasive malignancies. The PDI was 0.73 [95% CI: 0.70-0.76]. In case of random performance the PDI corresponds to 1/3.

Discrimination after updating

The average polytomous discrimination performance of the different updating methods on the 500 test sets is shown in Table 3. The c-statistic for the dichotomous models was identical for methods 0, 1 and 2 because the ranking of the predicted risks is maintained for recalibration (Table S3). However, when combining the dichotomous models to obtain a polytomous risk model, small decreases for the PDI were observed for methods 0 to 2. For the testicular cancer case study, revision (methods 3 and 4) or redevelopment resulted in a higher PDI on the updating set compared to no updating. This increase was not maintained in the test set, especially with an EPV of 3. For the ovarian tumor case study, redevelopment

showed similar discrimination as the original model on the updating set but lower discrimination on the test set.

Calibration after updating

Average calibration-in-the-large over the 500 test sets was poor when using the original prediction model, especially for the ovarian tumor case study (Table 4). Calibration was greatly improved by all other methods.

Updating parameters, variable selection and estimated effects

Intercept adjustments (method 1) were positive for the testicular cancer models, indicating that the original dichotomous models on average underestimated the probability of benign masses and the risk of teratomas (Table 5). The calibration slope (method 2) showed that the original model coefficients were too high in the benign vs malignant model (slope smaller than 1). For the teratoma vs viable cancer model the original coefficients were too low (slope larger than 1).

For the ovarian tumor case study, the positive intercept indicates a systematical underestimating of the probability of a benign tumor and of a borderline malignancy in the original dichotomous models. This may not be surprising as women presenting to public hospitals have much less aggressive tumors compared to women attending an oncology referral center. The slope adjustments were close to 1. The average recalibration results from method 2 were used to generate calibration plots of the original dichotomous models (Figure S1).

Method 3 selectively re-estimated individual coefficients. For the first model of the testicular cancer study all but one variable were frequently re-estimated despite the low number of

events per variable (Table S4). For the second model re-estimation was less frequent, related to similar effects of predictors and lower sample size. For the ovarian tumor study, re-estimation was rare, except for one predictor. Generally, effects that were often re-estimated with EPV 5 were less often re-estimated with EPV 3 due to a decrease in power. The reverse was seen for effects that were rarely re-estimated with EPV 5, due to an increase in variability.

Method 3 and 4 are revision methods that make use of shrinkage. The average shrinkage factors are shown in Table 5. If EPV 5 was associated with moderate to large shrinkage factors in revision methods 3 and 4, and as expected EPV 3 led to more shrinkage.

Model redevelopment (method 5), led to variable selection in agreement with the original models for the ovarian tumor study, and less so for the testicular cancer study (Table 6). This is consistent with the stronger slope adjustments by method 2 for the testicular cancer study, and the more frequent effect re-estimations by method 3, and the observation that the predictor effects differed more strongly between the original and the new setting for the testicular cancer study (Table S1, mainly the comparison of original coefficients with average coefficients for method 4). This explains why discrimination in the updating set could be increased by methods 3 to 5 for the testicular cancer study but not for the ovarian tumor study. Table 6 further demonstrated the instability of the variable selection results. The standard deviation of the number of selected variables increased when EPV decreased. Also, predictors with strong effects were selected less often and predictors with weak effects more often when EPV decreased. This is similar to the findings for method 3. Finally, the overfitting of model coefficients for model redevelopment can be seen in Table S1.

4. Discussion

Prediction models for polytomous outcomes can be very relevant for appropriate decision making, yet such outcomes often suffer from categories with low prevalence. We translated previously suggested methods to update existing dichotomous outcomes to the polytomous setting. The polytomous models were based on sequential dichotomous modeling. A model to predict category A vs category B+C was combined with a model to predict category B vs category C. Both dichotomous models were updated separately to obtain updated polytomous predictions. This approach was demonstrated on two case studies for which the updating setting involved other (types of) hospitals than the hospitals used for model development. Using the original model ('no updating') resulted in poor calibration whereas every other updating method greatly improved the calibration in a similar way. Regarding discrimination, we observed that none of the updating methods could improve performance compared to no updating. This is consistent with previous studies on model updating, and may be partly related to the insensitivity of c-statistics to detect differences in performance [22]. Overall, the findings in our study corroborate those obtained when updating dichotomous models [5,6,8,9,10,23].

Improvement of model discrimination through updating is more likely when predictor effects are strongly different in the new setting, and sample size of the updating data is sufficient. Further, again if sample size permits, model extension by adding one or more additional predictors may result in better discrimination depending on the situation. For example, a model can be extended with a newly available and promising marker, or with potentially predictive variables not considered when developing the original model [24]. Model extension was not considered in this study due to limited EPV. In the present study, as a

consequence of the low EPV, known perils such as model instability and overfitting also became visible in a polytomous setting when extensive updating methods were used. This tended to reduce performance in new patients. Not surprisingly, these problems were most pronounced for model redevelopment, with highly unstable variable selection results and overfitted model coefficients.

Sample size is typically a more severe problem for polytomous outcomes than for dichotomous outcomes, since one or more of the outcome categories often has very low prevalence. This necessitates the availability of large datasets to have a sufficient number of EPV for such categories, as clearly illustrated in our study. For the ovarian tumor case study the validation dataset contained 1107 patients, yet only 41 of them had a borderline malignant tumor. Therefore we varied the number of EPV even though this has already been studied extensively for dichotomous models. We simply divided the number of cases in the smallest category by the number of considered predictors, however for polytomous models more than one coefficient can be estimated per predictor. For example, we considered in the ovarian tumor case study seven predictors to predict a three-category outcome, such that two coefficients per predictor were estimated. If we adopt the common rule of thumb of EPV>10, we would need at least 7*2*10=140 borderline tumors. With a prevalence of 41 in 1107 (3.7%), a total sample size of at least 3800 patients is then recommended. For the testicular cancer study, a similar calculation yields a recommended sample size of at least 780 patients. Probably, one reason why polytomous outcomes are often dichotomized is to reduce sample size requirements.

More case studies are required to further support our findings. However, the findings are sufficiently interesting to focus future research on the extension of the updating methods to

prediction models based on multinomial logistic regression, which is the most common approach for polytomous outcomes. In addition, further research regarding calibration plots and the extension of recent performance measures such as integrated discrimination improvement and continuous net reclassification improvement is required [25,26].

Model validation is crucial before deploying a prediction model in routine clinical practice. This also includes an assessment of the differences in prevalence and patient characteristics between the patients from the setting where a model was developed and patients from the setting in which a model is planned to be used. This may reveal that updating would be advisable in order to ascertain that accurate probabilities are obtained. We conclude from this work that simple dichotomous updating methods also work well for models that predict polytomous outcomes. Recalibration can be easily done for sequential dichotomous models and will be the recommended approach when overall sample size for the updating data is small or when at least one of the outcome categories has a relatively low prevalence. However, if sufficient data are available, more complex updating methods such as revision, extension, or model redevelopment may become valid alternatives.

Acknowledgments

Kirsten Van Hoorde is supported by a PhD grant of the Flanders' Agency for Innovation by Science and Technology (IWT Vlaanderen). Ben Van Calster is a postdoctoral fellow from the Research Foundation – Flanders (FWO). Research supported by Research Council KUL (GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), PFV/10/002 (OPTEC)); Flemish Government (FWO project G.0493.12N, IWT project TBM070706-IOTA3, iMinds); and Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, `Dynamical systems, control and optimization', 2007-2011).

References

[1] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010; 21: 128–38

[2] Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine 2000; 19:453–73.

[3] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of Internal Medicine 1999; 130:515–24.

[4] König IR, Malley JD, Weimar C, Diener HC, Ziegler A. Practical experiences on the necessity of external validation. Statistics in Medicine 2007; 26:5499–511.

[5] Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statistics in Medicine 2004; 23:2567–86.

[6] Toll D, Janssen K, Vergouwe Y, Moons K. Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology 2008; 61:1085–94.

[7] Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. Statistics in Medicine 1995; 14:1999–2008.

[8] Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled?: Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999; 99:2098–104.

[9] Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. Journal of Clinical Epidemiology 2008; 61:76–86.

[10] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. United States of America: Springer; 2009.

[11] Frank E, Kramer S. Ensembles of nested dichotomies for multi-class problems. In Proceedings of the 21st International Conference on Machine Learning, Banff, Canada; 2004.

[12] Roukema J, van Loenhout RB, Steyerberg EW, Moons KG, Bleeker SE, Moll HA. Polytomous regression did not outperform dichotomous logistic regression in diagnosing serious bacterial infections in febrile children. Journal of Clinical Epidemiology 2008; 61:135–41.

[13] Biesheuvel C, Vergouwe Y, Steyerberg EW, Grobbee D, Moons K. Polytomous logistic regression analysis could be applied more often in diagnostic research. Journal of Clinical Epidemiology 2008; 61:125–34.

[14] Vergouwe Y, Steyerberg EW, Foster RS, Sleijfer DT, Fosså SD, Gerl A, de Wit R, Roberts JT, Habbema JDF. Predicting retroperitoneal histology in postchemotherapy testicular germ cell cancer: A model update and multicentre validation with more than 1000 patients. European Urology 2007; 51:424–32.

[15] Van Calster B, Valentin L, Van Holsbeke C, Testa AC, Bourne T, Van Huffel S, Timmerman D. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC Medical Research Methodology 2010; 10:96.

[16] Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. Journal of Clinical Oncology 2005; 23:8794–801.

[17] Van Holsbeke C, Van Calster B, Testa AC, Domali E, Lu C, Van Huffel S, Valentin L, Timmerman D. Prospective interval validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. Clinical Cancer Research 2009; 15:648–91.

[18] Timmerman D, Bourne T, De Rijdt S, Kaijser J, Van Calster B. Characterizing ovarian pathology: refining the performance of ultrasonography. International Journal of Gynecological Cancer 2012; 22:S9–S11.

[19] Steyerberg EW, Keizer HJ, Fosså SD, Sleijfer DT, Toner GC, Schraffordt Koops H, Mulders PF, Messemer JE, Ney K, Donohue JP. Prediction of residual retroperitoneal mass histology after chemotherapy for metastatic nonseminomatous germ cell tumor: multivariate analysis of individual patient data from six study groups. Journal of Clinical Oncology 1995; 13:1177–87.

[20] van Houwelingen JC, le Cessie S. Predictive value of statistical models. Statistics in Medicine 1990; 9: 1303–26.

[21] Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg E. Extending the c-statistic to nominal polytomous outcomes: The Polytomous Discrimination Index. Statistics in Medicine 2012; 31: 2610-26.

[22] Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Medical Research Methodology 2011, 11:13.

[23] Harrison DA, Brady AR, Parry GJ, Carpenter JR , Kathy R. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Critical Care Medicine 2006; 34:1378–88.

[24] Steyerberg EW, Pencina MJ, Lingsma HF, Kattan MW, Vickers AJ, Van Calster B. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. European Journal of Clinical Investigation 2012; 42:216–28.

[25] Pencina MJ, D' Agostino RB, D' Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008; 27:157–72.

[26] Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine 2011; 30:11–21.

Table 1: Description and descriptive statistics for the case studies.

Testicular cancer	**Development set (n=502)***	**Validation set (n=273)**
Setting	Different hospitals in three countries	Tertiary referral center from Indiana University
Outcome, n (%)		
Benign	237 (47%)	76 (28%)
Teratoma	213 (42%)	162 (59%)
Viable cancer	52 (10%)	35 (13%)
Variable, n (%) or mean (SD)		
Presence of teratoma elements	235 (46.8%)	104 (38.1%)
Elevated serum AFP	175 (34.9%)	68 (24.9%)
Elevated serum hCG	193 (38.4%)	75 (27.5%)
Square root (residual mass size)	4.75 (2.35)	7.79 (3.08)
Change in diameter (per 10%)	4.87 (5.73)	1.44 (5.73)

Ovarian tumor	**Development set (n=2037)**	**Validation set (n=1107)**
Setting	Tertiary referral centers for oncology in seven countries	Regional or public hospitals in five countries
Outcome, n(%)		
Benign	1315 (65%)	918 (83%)
Borderline	139 (7%)	41 (4%)
Invasive	583 (29%)	148 (13%)
Variable, n(%) or mean (SD)		
Age	48 (16)	46 (16)
Square root (ratio of solid and lesion sizes)	0.41 (0.42)	0.31 (0.37)
Masses on both ovaries	358 (17.6%)	214 (19.3%)
Ascites	297 (14.6%)	67 (6.1%)
Irregular cyst walls	923 (45.3%)	403 (36.4%)
Current use of hormonal therapy	269 (13.2%)	240 (21.7%)
Acoustic shadows	234 (11.5%)	156 (14.1%)

* 502 of 544 patients were selected by excluding 42 patients from the validation hospital (Indiana University) [14].

Table 2: Updating methods.

Method	Description
Method 0	the original prediction model without any adjustments
Method 1	the original prediction model where the intercept is adjusted
Method 2	the original prediction model where the intercept and the slope are adjusted
Method 3*	method 2 where it is tested whether the effect of the individual predictors is significantly different in the updating set; if not, the initially updated effect is kept
Method 4*	the original prediction model where the intercept and the regression coefficients of all predictors are re-estimated based on the data from the new setting
Method 5	building a new model from scratch for the updating set

* For methods 3 and 4 parameter shrinkage is used.

Table 3: Polytomous discrimination as assessed by the Polytomous Discrimination Index on the updating and test data. The values are averages over the 500 divisions of the validation data into updating and test datasets. The 5th and 95th percentiles of these results are shown between square brackets.

Testicular cancer		**Method 0**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Method 5**
EPV 5	Updating set	0.56 [0.52;0.60]	0.56 [0.52;0.60]	0.55 [0.51;0.59]	0.59 [0.55;0.63]	0.59 [0.55;0.64]	0.61 [0.57;0.65]
	Test set	0.57 [0.47;0.66]	0.56 [0.47;0.65]	0.55 [0.46;0.65]	0.57 [0.48;0.67]	0.58 [0.48;0.67]	0.58 [0.49;0.68]
EPV 3	Updating set	0.56 [0.50;0.63]	0.56 [0.50;0.63]	0.56 [0.49;0.63]	0.59 [0.52;0.67]	0.60 [0.52;0.68]	0.61 [0.52;0.70]
	Test set	0.56 [0.47;0.65]	0.56 [0.47;0.65]	0.55 [0.46;0.64]	0.56 [0.47;0.65]	0.57 [0.47;0.66]	0.55 [0.44;0.65]

Ovarian tumor		**Method 0**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Method 5**
EPV 5	Updating set	0.74 [0.72;0.75]	0.72 [0.71;0.74]	0.73 [0.71;0.74]	0.73 [0.71;0.75]	0.73 [0.71;0.75]	0.73 [0.71;0.76]
	Test set	0.73 [0.63;0.82]	0.72 [0.62;0.81]	0.72 [0.62;0.81]	0.72 [0.62;0.81]	0.72 [0.62;0.81]	0.70 [0.61;0.80]
EPV 3	Updating set	0.74 [0.69;0.78]	0.73 [0.68;0.77]	0.73 [0.68;0.78]	0.74 [0.69;0.79]	0.73 [0.69;0.78]	0.74 [0.70;0.79]
	Test set	0.73 [0.63;0.83]	0.72 [0.62;0.81]	0.72 [0.62;0.82]	0.72 [0.61;0.81]	0.72 [0.61;0.81]	0.70 [0.60;0.79]

Table 4: Calibration-in-the-large results on the test data, expressed as the difference in the observed prevalence (in %) and the expected prevalence (in %). The values are averages over the 500 divisions of the validation data into updating and test datasets. The 5th and 95th percentiles of these results are shown between square brackets.

Testicular cancer	**Outcome category**	**Method 0**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Method 5**
EPV 5	Benign	1.8 [-1.6;5.3]	0.1 [-4.9;5.0]	0.2 [-3.6;4.0]	0.2 [-3.9;4.3]	0.2 [-3.8;4.3]	0.2 [-4.3;4.5]
	Teratoma	1.8 [-1.5;4.5]	0.3 [-4.4;4.4]	0.4 [-3.7;3.9]	0.1 [-4.2;4.0]	0.0 [-4.4;3.9]	-0.2 [-4.3;3.8]
	Viable cancer	-3.6 [-5.1;-2.3]	-0.4 [-2.0;1.2]	-0.6 [-2.3;1.1]	-0.3 [-2.2;1.4]	-0.2 [-1.9;1.5]	0.0 [-2.3;2.0]
EPV 3	Benign	2.1 [-1.4;5.6]	0.0 [-5.6;5.5]	0.0 [-4.7;4.6]	0.0 [-5.0;4.8]	0.0 [-5.2;4.8]	0.1 [-5.5;5.2]
	Teratoma	1.6 [-1.8;4.6]	0.3 [-4.8;5.3]	0.5 [-4.2;4.5]	0.2 [-4.4;4.9]	0.2 [-4.7;4.9]	0.0 [-4.7;5.1]
	Viable cancer	-3.7 [-5.1;-2.2]	-0.3 [-2.0;1.4]	-0.5 [-2.5;1.6]	-0.3 [-2.4;1.9]	-0.2 [-2.1;2.0]	-0.1 [-2.9;2.5]

Ovarian tumor	**Outcome category**	**Method 0**	**Method 1**	**Method 2**	**Method 3**	**Method 4**	**Method 5**
EPV 5	Benign	7.2 [4.7;10.0]	-0.2 [-2.7;2.3]	-0.2 [-2.8;2.4]	-0.2 [-2.8;2.3]	-0.2 [-2.8;2.4]	-0.2 [-2.9;2.4]
	Borderline	-2.5 [-3.2;-1.8]	-0.4 [-1.1;0.2]	-0.3 [-1.0;0.3]	-0.2 [-0.9;0.4]	-0.3 [-1.0;0.4]	0.0 [-0.7;0.6]
	Invasive	-4.7 [-7.0;-2.4]	0.7 [-1.7;2.9]	0.5 [-1.9;2.8]	0.4 [-2.0;2.8]	0.5 [-1.9;2.8]	0.2 [-2.2;2.6]
EPV 3	Benign	7.4 [5.0;10.1]	0.0 [-2.7;2.6]	0.0 [-2.7;2.8]	0.1 [-2.6;2.9]	0.0 [-2.7;2.8]	0.1 [-2.8;2.8]
	Borderline	-2.5 [-3.2;-1.8]	-0.5 [-1.2;0.2]	-0.4 [-1.2;0.3]	-0.3 [-1.1;0.4]	-0.3 [-1.1;0.4]	-0.1 [-0.8;0.6]
	Invasive	-4.9 [-7.1;-2.7]	0.5 [-1.9;2.7]	0.3 [-2.2;2.8]	0.2 [-2.3;2.6]	0.3 [-2.2;2.6]	0.0 [-2.5;2.6]

Table 5: Average of the estimated updating parameters for the different updating methods.

Testicular cancer	**Benign vs malignant**		**Teratoma vs viable cancer**	
Method – parameter	**EPV 5**	**EPV 3**	**EPV 5**	**EPV 3**
Method 1 – intercept adj	0.12	0.14	0.26	0.26
Method 2 – intercept	-0.12	-0.08	0.08	0.06
Method 2 – slope	0.70	0.73	1.16	1.19
Method 3 – shrinkage	0.78	0.62	0.08	0.11
Method 4 – shrinkage	0.66	0.48	0.06	0.08

Ovarian tumor	**Benign vs malignant**		**Borderline vs invasive**	
Method – parameter	**EPV 5**	**EPV 3**	**EPV 5**	**EPV 3**
Method 1 – intercept adj	0.84	0.83	0.27	0.26
Method 2 – intercept	0.83	0.82	0.40	0.44
Method 2 – slope	1.10	1.11	1.11	1.16
Method 3 – shrinkage	0.35	0.24	0.29	0.23
Method 4 – shrinkage	0.06	0.08	0.12	0.13

Table 6: Frequency of variable selection and mean number of selected variables for model redevelopment (method 5) for the 500 updating sets.

Testicular cancer	**Benign vs malignant**		**Teratoma vs viable cancer**	
	EPV 5	**EPV 3**	**EPV 5**	**EPV 3**
Presence of teratoma elements	500	499	28	56
Elevated serum AFP	500	456	98	108
Elevated serum hCG	30	79	44	62
Square root (residual mass size)	16	48	499	448
Change in diameter	498	448	498	390
Mean number of selected variables (SD)	*3.1 (0.30)*	*3.1 (0.52)*	*2.3 (0.51)*	*2.1 (0.92)*

Ovarian tumors	**Benign vs malignant**		**Borderline vs invasive**	
	EPV 5	**EPV 3**	**EPV 5**	**EPV 3**
Age	500	500	500	500
Square root (ratio of solid and lesion sizes)	500	500	500	500
Masses on both ovaries	490	350	225	137
Ascites	500	500	195	163
Irregular cyst walls	500	499	1	20
Current use of hormonal therapy	316	208	1	31
Acoustic shadows	500	500	61	107
Mean number of selected variables (SD)	*6.6 (0.49)*	*6.1 (0.65)*	*3.0 (0.63)*	*2.9 (0.78)*

Table S1: Model coefficients for the original model, and average coefficients for the updating methods obtained on the 500 updating sets. For model redevelopment (method 5), only coefficients that differed from zero were considered to compute the average, i.e. coefficients from selected predictors.

		EPV 5				EPV 3			
Testicular cancer	**Original coefficients[a]**	**Method 2**	**Method 3**	**Method 4**	**Method 5**	**Method 2**	**Method 3**	**Method 4**	**Method 5**
Benign vs malignant									
Presence of teratoma elements	0.91	0.64	1.37	1.37	1.69	0.66	1.28	1.27	1.77
Elevated serum AFP	0.88	0.62	1.00	1.15	1.43	0.64	0.92	1.10	1.56
Elevated serum hCG	0.50	0.35	0.41	0.27	0.78	0.36	0.42	0.29	0.85
Square root (residual mass size)	-0.11	-0.08	-0.05	0.00	0.13	-0.08	-0.05	-0.01	0.09
Change in diameter	0.26	0.18	0.13	0.15	0.12	0.19	0.16	0.17	0.15
Teratoma vs viable									
Presence of teratoma elements	-0.65	-0.75	-0.71	-0.71	-0.91	-0.77	-0.70	-0.70	-0.91
Elevated serum AFP	-[b]	-[b]	-[b]	-[b]	-1.13	-[b]	-[b]	-[b]	1.08
Elevated serum hCG	-0.53	-0.62	-0.62	-0.60	-0.92	-0.64	-0.62	-0.59	-0.77
Square root (residual mass size)	-0.20	-0.23	-0.24	-0.24	-0.31	-0.24	-0.25	-0.25	-0.34
Change in diameter	-0.12	-0.13	-0.13	-0.14	-0.15	-0.14	-0.14	-0.14	-0.18

Ovarian tumors	**Original coefficients[a]**	**Method 2**	**Method 3**	**Method 4**	**Method 5**	**Method 2**	**Method 3**	**Method 4**	**Method 5**
Benign vs malignant									
Age	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03	-0.03	-0.04	-0.04
Square root (ratio of solid and lesion sizes)	-3.13	-3.44	-3.63	-3.47	-3.91	-3.47	-3.62	-3.53	-3.98
Masses on both ovaries	-0.59	-0.65	-0.63	-0.65	-0.64	-0.66	-0.65	-0.66	-0.79
Ascites	-2.24	-2.46	-2.39	-2.46	-2.42	-2.48	-2.45	-2.49	-2.50
Irregular cyst walls	-1.39	-1.53	-1.49	-1.52	-1.45	-1.54	-1.52	-1.54	-1.48
Current use of hormonal therapy	0.73	0.80	0.77	0.78	0.55	0.81	0.77	0.76	0.77
Acoustic shadows	2.13	2.34	2.18	2.29	1.71	2.36	2.21	2.29	1.81
Borderline vs Invasive									
Age	-0.03	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03	-0.04	-0.04
Square root (ratio of solid and lesion sizes)	-2.89	-3.22	-3.68	-3.45	-4.92	-3.35	-3.86	-3.71	-5.29
Masses on both ovaries	-0.72	-0.80	-0.72	-0.78	-0.93	-0.83	-0.77	-0.80	-1.40
Ascites	-1.65	-1.84	-1.61	-1.70	-1.19	-1.91	-1.67	-1.65	-3.60
Irregular cyst walls	0.45	0.50	0.45	0.43	1.04	0.52	0.45	0.42	-0.09
Current use of hormonal therapy	-[b]	-[b]	-[b]	-[b]	1.02	-[b]	-[b]	-[b]	-1.99
Acoustic shadows	-[b]	-[b]	-[b]	-[b]	-16.36	-[b]	-[b]	-[b]	-16.85

[a] The coefficients from the original model obtained in the development dataset. These coefficients are unchanged for updating methods 0 (no updating) and 1 (intercept adjustment).

[b] This variable was not considered in the original prediction model.

Table S2: Sizes of development set, updating set and test set.

Testicular cancer: total sample size (separate for benign - teratoma - viable cancer)			
EPV	**Development set**	**Updating set**	**Test set**
5	502 (237 - 213 -52)	195 (54 - 116 - 25)	78 (22 - 46 - 10)
3	502 (237 - 213 -52)	117 (33 - 69 - 15)	78 (22 - 46 - 10)

Ovarian tumor: total sample size (separate for benign - borderline - invasive)			
EPV	**Development set**	**Updating set**	**Test set**
5	2037 (1315 - 139 -583)	945 (784 - 35 - 126)	162 (134 - 6 - 22)
3	2037 (1315 - 139 -583)	567 (470 - 21 - 76)	162 (134 - 6 - 22)

Table S3: C-statistics for the sequential dichotomous models averaged over the 500 updating and test datasets. The 5th and 95th percentiles of these results are shown between square brackets.

	EPV 5						EPV 3					
Testicular cancer	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**
Benign vs malignant												
Updating set	0.78 [0.75; 0.81]	0.78 [0.75; 0.81]	0.78 [0.75; 0.81]	0.82 [0.78; 0.85]	0.82 [0.79; 0.85]	0.82 [0.79; 0.85]	0.78 [0.73; 0.84]	0.78 [0.73; 0.84]	0.78 [0.73; 0.84]	0.81 [0.75; 0.87]	0.82 [0.77; 0.88]	0.82 [0.76; 0.88]
Test set	0.78 [0.70; 0.86]	0.78 [0.70; 0.86]	0.78 [0.70; 0.86]	0.80 [0.72; 0.87]	0.81 [0.73; 0.88]	0.81 [0.73; 0.88]	0.78 [0.70; 0.86]	0.78 [0.70; 0.86]	0.78 [0.70; 0.83]	0.79 [0.70; 0.87]	0.79 [0.72; 0.87]	0.79 [0.70; 0.88]
Teratoma vs viable cancer												
Updating set	0.67 [0.62; 0.72]	0.67 [0.62; 0.72]	0.67 [0.62; 0.72]	0.68 [0.62; 0.73]	0.68 [0.62; 0.74]	0.70 [0.64; 0.75]	0.67 [0.58; 0.76]	0.67 [0.58; 0.76]	0.67 [0.58; 0.76]	0.69 [0.58; 0.78]	0.69 [0.59; 0.79]	0.71 [0.50; 0.81]
Test set	0.68 [0.54; 0.80]	0.68 [0.54; 0.80]	0.68 [0.54; 0.80]	0.67 [0.54; 0.80]	0.67 [0.54; 0.80]	0.66 [0.52; 0.80]	0.68 [0.54; 0.79]	0.68 [0.54; 0.79]	0.68 [0.54; 0.79]	0.67 [0.53; 0.79]	0.67 [0.53; 0.79]	0.63 [0.50; 0.77]
Ovarian cancer	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**
Benign vs malignant												
Updating set	0.92 [0.92; 0.93]	0.92 [0.92; 0.93]	0.92 [0.92; 0.93]	0.93 [0.92; 0.93]	0.92 [0.92; 0.93]	0.93 [0.92; 0.93]	0.92 [0.91; 0.94]	0.92 [0.91; 0.94]	0.92 [0.91; 0.94]	0.93 [0.91; 0.94]	0.93 [0.91; 0.94]	0.93 [0.91; 0.94]
Test set	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.95]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.88; 0.96]	0.92 [0.87; 0.95]
Borderline vs invasive												
Updating set	0.83 [0.81; 0.86]	0.83 [0.81; 0.86]	0.83 [0.81; 0.86]	0.84 [0.82; 0.87]	0.84 [0.81; 0.87]	0.85 [0.83; 0.88]	0.84 [0.78; 0.89]	0.84 [0.78; 0.89]	0.84 [0.78; 0.89]	0.85 [0.79; 0.90]	0.84 [0.79; 0.90]	0.86 [0.81; 0.91]
Test set	0.83 [0.69; 0.95]	0.83 [0.69; 0.95]	0.83 [0.69; 0.95]	0.83 [0.69; 0.95]	0.83 [0.69; 0.95]	0.77 [0.35; 0.93]	0.83 [0.70; 0.95]	0.83 [0.70; 0.95]	0.83 [0.70; 0.95]	0.83 [0.70; 0.95]	0.83 [0.70; 0.95]	0.79 [0.57; 0.95]

Table S4: Frequency of effect re-estimation for method 3 for the 500 updating sets.

Testicular cancer	**Benign vs malignant**		**Teratoma vs viable cancer**	
	EPV 5	**EPV 3**	**EPV 5**	**EPV 3**
Presence of teratoma elements	284	214	30	29
Elevated serum AFP	155	68	-[a]	-[a]
Elevated serum hCG	8	14	0	5
Square root (residual mass size)	124	90	36	52
Change in diameter	217	165	0	5

Ovarian tumors	**Benign vs malignant**		**Borderline vs invasive**	
	EPV 5	**EPV 3**	**EPV 5**	**EPV 3**
Age	0	3	0	4
Square root (ratio of solid and lesion sizes)	216	101	214	131
Masses on both ovaries	0	4	0	5
Ascites	0	0	29	35
Irregular cyst walls	0	8	1	16
Current use of hormonal therapy	8	18	-[a]	-[a]
Acoustic shadows	74	69	-[a]	-[a]

[a] This variable was not considered in the original prediction model and is therefore not considered for re-estimation.

Figure S1: Logistic calibration curves for the original sequential dichotomous models. The plots use the average intercept and slope adjustments for method 2 with EPV 5 over the 500 updating sets. Panel a gives the curves for the testicular cancer study, panel b for the ovarian tumor study.

(a)

(b)