Gynecologic Oncology Reports 18 (2016) 49–50

Contents lists available at ScienceDirect

Gynecologic Oncology Reports

journal homepage: www.elsevier.com/locate/gynor

Correspondence

Flawed external validation study of the ADNEX model to diagnose ovarian canceron behalf of TG6 of the STRATOS initiative

Dear Editor,

External validation studies of prediction models are of utmost importance in order to assess the performance of a prediction model in different locations (Altman et al., 2009). We therefore read with interest the recent external validation study of the ADNEX model (Szubert et al., 2016).

For patients with a persistent adnexal tumor who are scheduled for surgery, the ADNEX model predicts the risk of five tumor types: benign, borderline malignant, stage I cancer, stage II–IV cancer, or secondary metastatic cancer (Van Calster et al., 2014). The model was developed on data from 5909 patients collected at 24 centers, in 10 countries, between 1999 and 2012. ADNEX aims to assist clinicians make appropriate clinical decisions for patients presenting with an adnexal mass. When validating the ADNEX model, it is natural to first evaluate the prediction of malignancy, followed by the multiclass prediction of malignancy subtypes, in a similar way to other validation studies of multiclass models (Steyerberg et al., 1998). This approach is followed in the recent paper, but there are a number of important issues around the design, analysis, and reporting we wish to raise.

First, validation studies should be designed to reliably assess performance in terms of discrimination and calibration (Steyerberg, 2009). In this particular case, the authors report a sample size calculation for testing the hypothesis that the AUC of the model is higher than 0.5. Assuming an AUC of 0.94 leads to a very low required sample size ($n = 22$). This approach is at odds with methodological guidance and the result is that the precision of performance measures will be low: for dichotomous prediction, previous studies have suggested that at least 100, and preferably at least 200 individuals with the event (in this case ovarian malignancy) are required for a meaningful validation (Steyerberg, 2009; Vergouwe et al., 2005; Collins et al., 2016). Here, center 1 has 70 malignant tumors, whilst center 2 has only 34, leading to unreliable per center results. Validation would therefore best be done on all patients, with center-specific results as an exploratory addition. Furthermore, statistical tests to compare results between centers are provided throughout the text. Although heterogeneity of performance across locations is important (Riley et al., 2016), p-values to compare two specific centers are uninformative. It is useful to observe that the AUCs were 0.955 and 0.907, since this is in line with the center-specific values reported in the original publication describing the ADNEX model (Van Calster et al., 2014). A detailed investigation of heterogeneity should however involve a larger dataset with patients from many different centers. Furthermore, subgroup analyses by menopausal status become very unreliable when stratified by center.

Second, the authors have not adequately described their population and results. The prevalence of each of the five tumor types is not clearly provided, and the prevalence of stage I cancer and stage II–IV cancer can only be derived from the confusion matrix. The ADNEX model has variants with and without the serum marker CA125 as a predictor. The authors mix both variants depending on the availability of CA125, such that it is unclear to what variant the reported performance is referring.

Third, the calibration of the predicted risk of malignancy has not been investigated, i.e. whether observed frequencies of malignancy correspond to predicted risks, especially around the risk threshold of 10%. Unfortunately, this aspect of risk prediction models is often overlooked despite its importance (Steyerberg, 2009).

Finally, the 'multiclass' performance evaluation is fundamentally flawed. The key problem is the confusion matrix, which classifies patients into one of the five tumor types by choosing the group with the highest predicted risk. Baseline risk, or prevalence, of each tumor type varies substantially: among 327 patients, 223 are benign tumors (68%), 16 borderline (5%), 14 stage I primary cancers (4%), 64 stage II–IV primary cancers (20%), and 10 secondary metastatic cancers (3%). Given these large differences in prevalence, it is unlikely that ADNEX based risk predictions for secondary metastatic cancer will be larger than those for a benign tumor. As a result, the confusion matrix will rarely classify a tumor as a metastatic cancer, resulting in near zero sensitivity for this tumor type. Analogous arguments apply to borderline tumors and stage I primary cancers. Such results are misleading, since they are unrelated to the model's ability to discriminate between tumor types. More generally, it makes little clinical sense to classify patients into only one category. It is much more relevant to monitor which risks are high or increased, and to act upon them accordingly. For example, the predicted risk of advanced-stage ovarian cancer and the risk of secondary metastasis might both be increased (although the latter will usually be smaller than the former due to the lower prevalence). In such cases the clinician may focus management decisions on both tumor types. An elevated risk of a metastatic tumor may trigger planning additional preoperative diagnostic tests, such as gastroscopy, x-ray mammography or a full body MRI. Instead of a confusion matrix, concordance or *c* statistics for subgroup discrimination should be given. We would advise to present pairwise *c* statistics using the conditional risk method (Van Calster et al., 2012, 2014), although other approaches could be followed. Nevertheless, we warn that in this study the sample size is far too small to draw meaningful conclusions, although we realize that it would require a very large sample to have information on 100 secondary metastatic cancers, as in the IOTA collaboration (Van Calster et al., 2014).

In conclusion, we are happy to observe the excellent discrimination between benign and malignant tumors seen in this study, in line with the original publication (Van Calster et al., 2014). However, the analysis does not allow us to draw any reliable conclusions with respect to multiclass discrimination. To improve reporting of prediction model

DOIs of original article: http://dx.doi.org/10.1016/j.gore.2016.10.009, http://dx.doi.org/10.1016/j.ygyno.2016.06.020

http://dx.doi.org/10.1016/j.gore.2016.09.003

studies, the TRIPOD guidelines have recently been introduced (Moons et al., 2015). These guidelines highlight the need for adequate sample size, assessment of calibration and transparent reporting of key information such as number of events in each category. Although we recognize that validation of multiclass models involves additional difficulties, it is clear that the TRIPOD recommendations should be followed to ensure all key information is clearly reported.

Conflict of interest

The authors declare no conflicts of interest.

References

Altman, D.G., Vergouwe, Y., Royston, P., Moons, K.G., 2009. Prognosis and prognostic research: validating a prognostic model. BMJ 338, b605.

Collins, G.S., Ogundimu, E.O., Altman, D.G., 2016. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat. Med. 35, 214–226.

Moons, K.G., Altman, D.G., Reitsma, J.B., Ioannidis, J.P., Macaskill, P., Steyerberg, E.W., et al., 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med. 162, W1–73.

Riley, R.D., Ensor, J., Snell, K.I., Debray, T.P., Altman, D.G., Moons, K.G., et al., 2016. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 353, i3140.

Steyerberg, E.W., 2009. Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating. Springer-Verlag, New York.

Steyerberg, E.W., Gerl, A., Fossa, S.D., Sleijfer, D.T., de Wit, R., Kirkels, W.J., et al., 1998. Validity of predictions of residual retroperitoneal mass histology in nonseminomatous testicular cancer. J. Clin. Oncol. 16, 269–274.

Szubert, S., Wojtowicz, A., Moszynski, R., Zywica, P., Dyczkowski, K., Stachowiak, A., et al., 2016. External validation of the IOTA ADNEX model performed by two independent gynecologic centers. Gynecol. Oncol.

Van Calster, B., Vergouwe, Y., Looman, C.W., Van Belle, V., Timmerman, D., Steyerberg, E.W., 2012. Assessing the discriminative ability of risk models for more than two outcome categories. Eur. J. Epidemiol. 27, 761–770.

Van Calster, B., Van Hoorde, K., Valentin, L., Testa, A.C., Fischerova, D., Van Holsbeke, C., et al., 2014. 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. BMJ 349, g5920.

Vergouwe, Y., Steyerberg, E.W., Eijkemans, M.J., Habbema, J.D., 2005. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J. Clin. Epidemiol. 58, 475–483.

B Van Calster[1]
KU Leuven, Department of Development and Regeneration, Leuven, Belgium
Department of Public Health, Erasmus MC, Rotterdam, The Netherlands
Corresponding author at: KU Leuven, Department of Development and Regeneration, Leuven, Belgium.
E-mail address: ben.vancalster@kuleuven.be.

EW Steyerberg[1]
Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

T Bourne
KU Leuven, Department of Development and Regeneration, Leuven, Belgium
Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium
Queen Charlotte's & Chelsea Hospital, Imperial College London, London, UK

D Timmerman
KU Leuven, Department of Development and Regeneration, Leuven, Belgium
Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium

GS Collins[1]
Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

24 July 2016

[1] STRATOS initiative (STRengthening Analytical Thinking for Observational Studies), Topic Group 6 on evaluating diagnostic tests and prediction models.