**Multinomial risk models for ovarian cancer diagnosis:**

**Comparison of algorithms**

Statistical Analysis Plan

January 2021

Updated June 2021 regarding multiple imputation approach

Ben Van Calster, KU Leuven

1. **Overall aim and research questions**

The currently best performing prediction model for ovarian cancer diagnosis is the ADNEX model. ADNEX is a multinomial logistic regression model that can be used to estimate the risk that an adnexal tumor is benign, borderline, stage I primary invasive, stage II-IV primary invasive or a secondary metastasis.1 It was developed on data from 5909 patients with an adnexal tumor that was operated on within 90/120 days of the decision to operate.1 Patients were recruited at 24 international centers between 1999 and 2012. ADNEX has a version with CA125 as a predictor, and a version without CA125. ADNEX has been externally validated in a large multicenter study involving 4905 patients recruited at 17 international centers.2 The AUC for benign versus malignant was 0.94 for the versions with and without CA125. Of note, this external validation study included all women with an adnexal tumor that was seen for the first time in a center, irrespective of whether the tumor was operated on or followed conservatively. 2489 of the 4905 patients were operated without follow-up scan within 120 days, and are therefore comparable to the 5909 patients on which ADNEX was developed. On these 2489 patients, ADNEX had an AUC of 0.92 when CA125 was included and 0.91 when CA125 was excluded.

Currently, there is strong interest in other more flexible algorithms for developing predictive models.3 Although on average, such models may not add much to standard regression-based models if the study was properly conducted, the relative merit of different algorithms may depend on the context.4 Earlier effects to compare flexible models with regression models for ovarian cancer diagnosis did not reveal better predictive ability, but these efforts were based on kernel-based algorithms and neural networks, small datasets, and incomplete performance assessments.5,6

Therefore, the aims of this study are

1. Use the 5909 patients on which ADNEX was developed to develop models based on different types of algorithms for binary (benign vs malignant) and multiclass (5 categories) risk prediction
2. Validate these models on the 2489 validation patients who were operated within 120 days without follow-up scan, and compare performance in terms of discrimination, calibration and clinical utility
3. **Data source**

The data to develop the models is the same data that was used to develop the ADNEX model. These are pooled data from the IOTA study phases 1, 1b, 2, and 3. The data to validate the model are data from the interim analysis of IOTA5 that have been used to validate ADNEX and related models. The primary analysis involved data from 4905 patients recruited at 17 centers between 2012 and 2015. Of these 4905 patients, 2489 were operated without follow-up ultrasound examination within 120 days of the inclusion scan.

1. **Tumor outcome (reference standard)**

For the 5909 model development patients, the reference standard was based on the histological examination of excised tumor tissue. Based on the histological characterization, tumors were classified as benign, borderline, stage I primary invasive, stage II primary invasive, or secondary metastasis. For the validation patients, the same procedure was followed for the 2489 patients with surgery within 120 days.

1. **Predictor variables and parameters**

ADNEX uses 9 predictor variables. Six variables are ultrasound variables: maximal diameter of the lesion, proportion of solid tissue, number of papillary projections, presence of more than 10 cyst locules, acoustic shadows, and ascites in the pouch of Douglas. The other predictors are age, CA125 level, and type of center. Type of center refers to whether the center is a tertiary oncological referral center or not. Of note, 10 predictors were considered when developing ADNEX, but family history of ovarian cancer was excluded.

Additional predictors could be: family history of ovarian cancer (yes v no), color score of tumor vascularity (ordinal, 4 levels), bilateral tumors (yes v no), personal history of ovarian cancer (yes v no), and echogenicity (categorical, 6 levels), papillations with blood flow (yes v no), irregular internal cyst walls (yes v no).

When developing ADNEX, transformations for continuous variables were investigated. As a result, age was used linearly, CA125 and maximal diameter of the lesion were log-transformed, for proportion solid tissue a linear and quadratic term were used, and number of papillary projections was used linearly (i.e. variable is used as a numerical variable). This leads to 10 parameters, although the development procedure had 17 degrees of freedom in order to select these 10 parameters. But note that the development procedure was performed on 3506 of 5909 patients, and hence 2403 patients were not used at all for parameter selection.

1. **Sample size**

The least common outcome category is malignancy for the binary outcome, and secondary metastasis for the multiclass outcome. The development dataset has 3980 benign tumors, 1929 malignancies and 246 secondary metastases. With 10 parameters, this means 1929/10=192 malignancies and 246/10=24 secondary metasases per parameter. However, the 10 parameters would require 40 coefficients. In that sense, there are 1929/40=48 malignancies and 246/48=6 secondary metasases per coefficient.7 This is low for the multinomial outcome, although lower values are acceptable when some outcome categories are rare and others are very common as is the case here.7 E.g. there are 3980 benign tumors in the development dataset, which is 3980/40=99 per coefficient. The ADNEX model included ‘uniform shrinkage’ of model coefficients,1,8 and the recent validation study did not provide evidence of overfitted (overly extreme) risk estimates.2 Flexible algorithms may be much more data hungry than regression algorithms, however, which would raise further concerns for multinomial model development. In the end, we will evaluate the developed models on an independent external validation dataset, which will allow us to study model performance.

The validation dataset has 132 observed secondary metastases. This size of the least common category is commonly seen as acceptable.9,10

1. **Modeling strategy**
   1. Prediction models

We will fit the following models for the binary outcome (benign v malignant) using the ADNEX variables without CA125:

* Logistic regression with the same variable transformations in ADNEX
* Logistic regression without transformations
* Ridge logistic regression with the ADNEX transformations
* Firth logistic regression with the ADNEX transformations
* Support vector machine with Gaussian kernel (no transformations)
* Random forest (no transformations)
* Extreme gradient boosting (Xgboost) (no transformations)
* CART (no transformations)
* Neural network (multilayer perceptron with 1 hidden layer) (no transformations)

We will fit the following models for the multiclass outcome using the ADNEX variables without CA125:

* Multinomial logistic regression with the same variable transformations in ADNEX
* Multinomial logistic regression without transformations
* Multinomial ridge logistic regression with the ADNEX transformations
* Multinomial Firth logistic regression with the ADNEX transformations
* Support vector machine with Gaussian kernel (no transformations); for multiclass outcomes, one may perform binary SVMs for each pair of outcomes, and combine the resulting probabilities with pairwise coupling
* Random forest (no transformations)
* Extreme gradient boosting (Xgboost) (no transformations)
* CART (no transformations)
* Neural network (multilayer perceptron with 1 hidden layer) (no transformations)

This can be repeated with CA125 as a predictor. Then, multiple imputation for missing values has to be used. 100 multiple imputations already exist for the development and the validation data. We will develop models using the first 10 imputations, which means that 10 models are developed for each algorithm. These 10 models are then applied to each of 10 imputed validation datasets. For each imputed validation dataset, the predictions from the 10 models are averaged. Using the average predictions, model performance can be quantified.

Then, these models can also be developed after adding the abovementioned list of additional predictors.

Ridge models have one hyperparameter (lambda), SVMs have two (sigma of Gaussian kernel and the regularization/cost parameter), for random forests we will tune two (mtry and minimal node size), for CART with tune two (minimum cases in a leaf node, and minimum cases in a node to attempt a split), for neural networks we tune two (number of hidden neurons and decay), and for xgboost we tune two (learning rate / eta, and booster method).11,12 Tuning will be done using 10-fold cross-validation.

* 1. Analysis population

The primary analysis population involves the women with an adnexal mass that was operated within 120 days. If a patient has multiple masses, the dominant mass is used. For models that include CA125, multiply imputed data are used (and available) to handle missing CA125 levels.

* 1. Model development and validation procedures
     1. Final models and ‘apparent’ performance

Models are developed in R, using for example the glmnet package for ridge, the brglm2 package for Firth logistic regression, and caret for the machine learning algorithms.

For models that include CA125 and therefore use multiply imputed data, the final model is obtained as follows: a model is fitted on each imputed dataset, leading to 10 models. All models are then applied to each imputed dataset, and predictions are averaged in each imputed dataset. Using these averaged predictions, performance estimates are obtained in each imputed dataset. These estimates are combined using Rubin’s rules17 to get final apparent performance estimates.

Apparent performance, i.e. performance on the very same data on which the model was developed, will be quantified using the AUC for benign vs malignant, the Polytomous Discrimination Index (PDI) (multiclass models only), the c-statistic for every pair of outcome categories using the conditional risk method (multiclass models only).13,14 Apparent calibration performance is visualized using (multinomial) calibration curves.15

* + 1. External validation

The external validation will mostly be the same as in the recently published large multicenter validation study of the ADNEX model.2

* AUC benign versus malignant: this is calculated per center, and then center-specific results are combined using meta-analysis. The overall estimated risk of malignancy is 1 minus the estimated risk of benign, which equals the sum of the estimated risks of the 4 types of malignancy. Calculations are done using logit(AUC) and its standard error, and are transformed back in the end. The auc.nonpara.mw function of the auRoc R package can calculate this. Otherwise meta-analysis packages such as metafor may allow the specification of the logit transformation. Results can be presented as forest plots.
* (Multiclass models only) The Polytomous Discrimination Index as a multinomial extension of the AUC: this is calculated per center, and then combined using meta-analysis. The mcca R package gives PDI and its standard error. Perhaps a logit transformation would be of interest here too, but there is no research on this.
* (Multiclass models only) Pairwise AUC for every pair of outcome categories using the conditional risk method. The AUC for categories A and B is based on the conditional risk . This is based on pooled data due to small n in some outcome categories when results are calculated per center. Calculations can be based on logit(pairwise AUC).
* Calibration slope for benign versus malignancy:16 center-specific and overall results can be obtained in one step, by fitting a logistic regression model of binary outcome (1, malignant; 0, benign) on the logit of the overall estimated risk of malignancy with a random intercept and slope by center. The fixed effect of logit(risk of malignancy) is the overall slope. Center-specific results can be obtained using the random slopes per center, but need not be shown.
* Calibration intercept for benign versus malignancy:16 this is based on a similar model as in the previous point, but now the coefficient for logit(risk of malignancy) is fixed to 1. This can be done by adding this variable as an offset. The fixed intercept yields the overall calibration intercept.
* Calibration curves:16 center-specific and overall curves can be based on the results of the model used to obtain calibration slopes. The overall curve is based on the fixed effects, center-specific curves take random intercept and slope per center into account.
* (Multiclass models only) Flexible multinomial calibration curves:15 these are obtained by using a pooled (i.e. fixed effects only) multinomial calibration model. This is a multinomial logistic regression of the outcome (5 levels) on the 4 linear predictors.

Notice that for center-specific results, the four smallest centers will be combined (London, IUK; Milan 3, MIT; Florence, FLI; Nottingham, NUK).2

When validation with multiple imputed data is performed, the following needs to be done. Per center, AUC, PDI and pairwise AUCs are first combined across the imputed datasets using Rubin’s rules.17 The calibration slope and calibration intercept models are fitted in each imputed dataset (based on averaged predictions). Then, the fixed effects are combined using Rubin’s rules, and the random effects are simply averaged. For multinomial calibration curves, the analysis is done per imputed dataset, and the curves are averaged. Multiple imputation in validation data is an issue for models including CA125.

* + 1. Statistical tools and software
* Model fitting: glmnet, brglm2, caret
* Calculation of standard c-statistic (area under the receiver operating characteristic using estimated risk of EP/PPUL, e.g. using auRoc package in R and auc.nonpara.mw function with method=”pepe”). To use Rubin rules, it is better to do that using logit(c) and the standard error for the logit(c). In the end, the inverse logit can be used to backtransform to the original scale.
* Calculation of Polytomous Discrimination Index, mcca package in R.7,12
* Calculation of pairwise c-statistics using conditional risk method.8
* Combining results across imputed datasets using the Rubin rules.17 See e.g. MIcombine of mitools R package)

1. **References**
2. Van Calster B, et al. BMJ 2014.
3. Van Calster B, et al. BMJ 2020.
4. Goldstein BA, et al. Eur Heart J 2017.
5. Christodoulou E, et al. J Clin Epidemiol 2019.
6. Van Holsbeke C, et al. Clin Cancer Res 2012.
7. Van Calster B, et al. BMC Med Res Methodol 2010.
8. De Jong V, et al. Stat Med 2019.
9. Van Calster B, et al. Stat Methods Med Res 2020.
10. Vergouwe Y, et al. J Clin Epidemiol 2005.
11. Collins GS, et al. Stat Med 2016.
12. Kuhn M, Johnson K. Applied Predictive Modeling. Springer 2013.
13. Probst P, et al. J Mach Learn Res 2019.
14. Van Calster B, et al. Stat Med 2012;31:2610-2626.
15. Van Calster B, et al. Eur J Epidemiol 2012;27:761-770.
16. Van Hoorde K, et al. Stat Med 2014.
17. Van Calster B, et al. BMC Med 2019.
18. Sterne JAC, et al. BMJ 2009;338:b2393.