IOTA5 interim analysis

Statistical Analysis Plan

Page 2-11: Follow-up study (27 December 2017)

Page 12-19: Performance of models for ovarian tumor diagnosis when applied to all patients (operated or conservatively followed) (28 May 2019)

IOTA5 interim analysis

Follow-up study

Statistical Analysis Plan

27 December 2017

1.1 Data:

The IOTA5 database will be used for the analysis and the following in- and exclusion criteria will be used.

Inclusion criteria:

- Initial policy is conservative management,
- The mass was diagnosed as benign at the first visit (i.e. subjective assessment = benign AND certainty = certainly benign OR probably benign).
- First scan between 1 January 2012 and 1 March 2015.

Exclusion criteria:

- Patients who did not give consent; data from patients who withdrew consent during the study will not be used either
- Younger than 18 years at first scan
- Cysts that are deemed clearly physiological and less than 3 cm in maximum diameter are not eligible for inclusion. *[We identified 13 such patients in the database for the interim analysis.]*

Further issues regarding data:

- Patients who underwent surgery for the adnexal mass after the first scan, despite the initial policy being conservative management, will be used in descriptive statistics but not in the main survival analysis. We will include them in a sensitivity analysis (cf below).
- Patients who withdrew consent during the study will not be used
- Only the dominant mass of the patient will be included
- We include follow-up data up to 30th of June 2017 such that all patients should have been followed up for 24 months (protocol allows +/- 2 months, hence this date allows 26 months follow-up for all patients; for the survival analysis, we use all available data (i.e. also information on the first scan after 26 months if available) but only show results up to 30 months. 30 months implies 2.5 years, this is convenient when constructing plots. Follow-up after 30 months is incomplete and hence results are uncertain.
- Although the aim of IOTA5 was to recruit patients consecutively and organize follow-up, there is a lot of heterogeneity among centers regarding adherence to the follow-up protocol. Therefore, we will perform a sensitivity analysis including all centers that have contributed to patient recruitment. However, only data of the following centers will be used for the primary analysis (centers with sufficient data quality: based on N recruited for conservative

management and % useful info): AGR, CIT, DEP, FLI, GBE, IUK, LBE, MIT, MPO, MSW, NCI, NUK, OIT, PSP, RIT, SIT, SSW, TIT and TUS.

1.2 Analysis:

1.2.1 Population

The population in this study consists of cases with a new adnexal mass (not already in follow-up in the center), and of patients with a mass that was already in follow-up in the center before the center started recruitment for the IOTA5 study. The latter group is likely a selected one: known masses that have spontaneously resolved or were operated on before the center started recruiting for the IOTA5 study will not be included. In addition, follow-up of these masses before the start of IOTA5 is ignored. Therefore, we assume that the patients with a known mass already in follow-up will more likely have a persistent and harmless mass. In this group, we can expect fewer spontaneous resolutions, less surgery for suspicion of malignancy or other complications, resulting in fewer confirmed complications. First, we will analyze the follow-up for all patients (new + known masses) in order to describe fully what was found during the IOTA5 study. Then, we will focus on follow-up results for patients with new masses only in order to obtain statistically more correct estimates.

Patients that have been selected for conservative management after subjective assessment by the ultrasound investigator, can nevertheless undergo surgery after inclusion, before having proceeded to follow-up (because of many reasons e.g. decision of the managing clinician, patient request etcetera). In the primary analysis we will describe this group with immediate surgery (why surgery was performed, what was found at surgery), but for the survival analysis only cases with at least one follow-up scan will be included (intention to treat with follow-up). In the appendix, we will add a sensitivity analysis that involves a survival analysis including patients with immediate surgery (intention to treat with or without follow-up).

1.2.2 Methods

To describe findings during follow up, we will use cumulative incidence curves that reflect the competing risk setting. The follow-up time will be computed in months.

In competing risk survival analysis, the hazard and cumulative incidence of competing (i.e. mutually exclusive) events are estimated over time. The basic competing risk outcome levels reflect the overall study outcome, and are ‘surgery performed’, ‘cyst spontaneously resolved’, and ‘patient died (for any reason)’, whichever comes first. When surgery is performed, this can be further subdivided based on reason for surgery and findings at surgery. For these subdivisions, it is possible that two events (i.e. reasons or findings) occur simultaneously. For example, a patient can undergo surgery due to acute pain and fertility concerns. We therefore specified a hierarchy of the competing events and if two events occur simultaneously, the event highest in the hierarchy is given to the observation. The hierarchies are shown in Box 1. For example, when the patient is operated due to acute pain and fertility concerns, the patient is said to have the event acute pain.

Box 1. Hierarchy of reasons for surgery and findings at surgery.

Reasons for surgery	**Findings at surgery**
1. Suspicion of malignancy[a]	1. Invasive
2. Pain[b]	2. Borderline
3. Patient request/opportunistic[c]	3. Torsion
	4. Cyst rupture
	5. Minor complications[d]
	6. No complications of tumor

[a] Includes 'suspicion of malignancy based on ultrasound; suspicion of malignancy based on: increase in size, change in morphology, change in vascularity, raised or increase in CA125, raised HE4, CT finding, MRI findings'
[b] Includes 'acute pain (suspected torsion, suspected cyst rupture), chronic pain'
[c] Includes 'fertility concerns, patient request, opportunistic removal'
[d] Includes 'tumor bleeding, adhesions, inflammation/infection

We will derive cumulative incidence functions for study outcome, reason for surgery, and findings at surgery. For each outcome, figures depicting the cumulative incidence will be produced, as well as tables at time points 12 and 24 months containing the cumulative incidences of the competing events with 95% CI.

Median follow-up will be estimated by the reverse Kaplan-Meier method, i.e. where censoring is denoted as the event, and events denoted as censoring. This is the common method to visualize the censoring function.

Subgroup analyses will also be performed in the same manner as the overall analysis. Subgroup variables are tumour type (unilocular, unilocular-solid, multilocular, multilocular-solid, solid), presumed diagnosis (see table below), subjective assessment (certainly benign, probably benign), Lesion size (median split method + description on the cumulative incidence at cut-offs 3, 4, 5, 7, 10 cm), and menopausal status. For menopausal status, the data collection software contains a bug such that this information is overwritten by the menopausal status of the last visit. Therefore, for a small group of patients who are labeled postmenopausal, it is possible that they were premenopausal at the first visit in the study. However, we will use the variable as currently available in the dataset. If menopausal status is uncertain, we will classify it as postmenopausal if age at inclusion is 50 or higher.

In all subgroup analyses, we will report cumulative incidences on: spontaneous resolution, death (any cause), surgery (any reason), malignancy, torsion, rupture.

In the subgroup analyses, cumulative incidences will only be reported at 24 months.

Table. Presumed diagnoses at first visit

Presumed diagnosis	**N (%)**
Simple cyst / para-ovarian or salpingeal cyst/ Serous cystadenoma / serous cystadenofibroma	1470 (47%)
Endometrioma	609 (19%)
Teratoma	325 (10%)
Functional cyst	186 (6%)
Fibroma / fibrothecoma	180 (6%)
Mucinous cystadenoma / mucinous cystadenofibroma	129 (4%)
Hydrosalpinx	122 (4%)

1.2.3 Planned analyses, figures, tables

This list does not indicate which figures or tables will be included in the main document.

Figure: Flowchart of patients

Table: Patient and tumor characteristics at inclusion for patients with initial policy conservative (including patients that were operated immediately). *The list of variables to be described can be augmented with other relevant variables according to clinicians, and these will be added irrespective of the results.*

	All patients	Patients with newly detected masses	Patients with masses already in follow-up
Age			
Menopausal status			
Time in follow-up before inclusion		NA	
Tumor type			
Lesion size			
Solid parts			
…			
Presumed diagnosis (endometioma etc)			
…			
Subjective assessment (certainty)			
Immediately operated			
Patients with a newly detected masses, n (%)		NA	NA
Symptomatic at first scan			
Pain			
Abnormal bleeding			
Other symptoms			
Outcome events			
Cyst spontaneously resolved			
Surgery performed			
Invasive tumor			
Borderline tumor			
Cyst torsion			
Cyst rupture			
Complications detected at surgery*			
No complication detected at surgery			
Patient died			

Table: List of centers, with number of patients per center. This table will include all centers (all the ones included in the sensitivity analysis). Order by total number of patients (conservative and surgery). This table is intended to be included in the appendix or supplementary material.

Center	N	Initial policy surgery	Initial policy conservative	% No info	% Useful info
Malmo, Sweden					
Leuven, Belgium					
Rome, Italy					
…					
All oncology centers					
All non-oncology centers					
All centers					

Useful information: study outcome observed, or follow-up of at least 10 months.

Table: List of centers, with number of patients per center. This table will include centers with sufficient data quality (all the ones included in the primary analysis). Order by total number of patients (conservative). This table is intended to be included in the main report.

Center	Initial policy conservative	% Useful info
Malmo, Sweden		
Leuven, Belgium		
Rome, Italy		
…		
All oncology centers		
All non-oncology centers		
All centers		

Figure: Histogram of time of follow-up before recruitment in IOTA5 for patients with a mass already in follow-up at inclusion (‘old’ patients)

Figure: Histogram of time of follow up scans (all follow up scans, i.e. not only the last scan per patient). This will be done once based on data from all patients, and once based on data from patients with newly detected masses only.

Figure: Cumulative incidence curve for study outcome (died, resolution, surgery). This will be done once based on data from all patients, and once based on data from patients with newly detected masses only.

Figure: Cumulative incidence curves for reason for surgery (study outcome surgery is subdivided into different reasons as specified above in Box 1; only curves for these reasons are given, not for other study outcomes (died/resolved)). This will be done once based on data from all patients, and once based on data from patients with newly detected masses only.

Figure: Cumulative incidence curves for surgery outcome (study outcome is again subdivided, but now in terms of findings at surgery as specified above in Box 1). Panel A gives all events with y-axis from 0 to 1; Panel B given zoomed in plot focusing only on malignancy, torsion, rupture. This will be done once based on data from all patients, and once based on data from patients with newly detected masses only.

Table: Cumulative incidence of study outcomes, reasons for surgery, and findings at surgery at 12 and 24 months with 95% CI; Table containing the following statistics at time points 12 months and 24 months: cumulative incidence with standard error and 95% CI. This will be done once based on data from all patients, and once based on data from patients with newly detected masses only.

	12 months	24 months
Study outcome		
Cyst resolution		
Patient died		
Surgery performed		
Reason for surgery		
Suspicion of malignancy		
Pain[b]		
Patient request/opportunistic[c]		
Findings at surgery		
Invasive		
Borderline		
Torsion		
Cyst rupture		
Minor complications		
No complications		

Table: Table of all cases with confirmed malignancy, rupture, or torsion, with descriptive information. *The exact list of parameters to be described should still be defined.*

Case	New or known mass at inclusion	Time in follow up before surgery	Finding(s) at surgery	Age at inclusion	Lesion size at first scan	Change in lesion size	…
1							
2							
3							
4							
5							
6							
7							
…							

Table: Cumulative incidence of each outcome (death (any cause), spontaneous resolution, surgery (any reason), malignancy, torsion, rupture at 24 months with 95% CI for all levels of all subgroup variables for patients with new masses. This will be done based on data from patients with newly detected masses only.

Subgroup variable	Death (any cause)	Spontaneous resolution	Surgery (any reason)	Malignancy	Torsion	Rupture
Tumor type						
Unilocular						
Unilocular-solid						
…						
Presumed diagnosis						
Endometriosis						
…						
Subjective assessment						
Certainly benign						
Probably benign						
Lesion size						
median split (4 cm?)						
cut-offs 3, 4, 5, 7, 10 cm						
Menopausal status						
Premenopausal						
Postmenopausal						

Table: For cases with initial policy conservative but who were nevertheless immediately operated: overview of histology at surgery. This will be done based on data from patients with newly detected masses only.

Histology	N (%)
Endometrioma	
Dermoid	
...	
Borderline	
…	

Table: Overview of reasons for surgery for cases with initial policy conservative, stratified by whether they were nevertheless immediately operate d or not. This will be done based on data from patients with newly detected masses only.

Reason for surgery	Immediately operated N (%)	Patients with FU N (%)
Suspicion of malignancy		
Pain		
Suspected torsion		
Suspected cyst rupture		
Acute pain		
Chronic pain		
Patient request / opportunistic		
Fertility concerns		
Patient request, other		
Opportunistic		

Table. Sensitivity analysis using data from all centres. This will be done based on data from patients with newly detected masses only.

	12 months	24 months
Study outcome		
Cyst resolution		
Patient died		
Surgery performed		
Reason for surgery		
Suspicion of malignancy		
Pain[b]		
Patient request/opportunistic[c]		
Findings at surgery		
Invasive		
Borderline		
Torsion		
Cyst rupture		
Minor complications		
No complications		

Table. Sensitivity analysis including patients that were immediately operated. This will be done based on data from patients with newly detected masses only.

	12 months	24 months
Study outcome		
Cyst resolution		
Patient died		
Surgery performed		
Reason for surgery		
Suspicion of malignancy		
Pain[b]		
Patient request/opportunistic[c]		
Findings at surgery		
Invasive		
Borderline		
Torsion		
Cyst rupture		
Minor complications		
No complications		

IOTA5 interim analysis

Performance of models for ovarian tumor diagnosis when applied to all patients (operated or conservatively followed)

Statistical Analysis Plan

March 28th, 2019

(minor edits on May 28th, 2019)

1.1 Objective

To assess the performance of prediction models and subjective assessment for ovarian tumor diagnosis when evaluated on all patients presenting with an adnexal mass: patients operated after the recruitment scan (without having received conservative follow-up) and patients managed conservatively with ultrasound follow-up. Because this is not the same population as the population on which the models were developed (masses when selected for surgery), this is not a true validation study.

1.2 Population

Adult patients with an adnexal mass, who are included in the IOTA5 interim dataset (cf inclusion and exclusion criteria), except masses that are deemed clearly physiological if less than 3cm in maximal diameter. The interim dataset contains patients recruited in the IOTA5 study between 1 January 2012 and 1 March 2015, with follow-up data up to 30 June 2017.

1.3 Data

The IOTA5 database will be used for the analysis and the following in- and exclusion criteria will be used.

Inclusion criteria:

- Recruitment scan between 1 January 2012 and 1 March 2015.
- Patients who were not already in follow-up at a center before the recruitment scan ('new' patients) (i.e. new patients)
- Patients recruited in centers AGR (Athens), CIT (Milan), FLI (Florence), GBE (Genk), IUK (London), LBE (Leuven), MIT (Milan), MPO (Katowice), MSW (Malmö), NCI (Milan), NUK (Nottingham), OIT (Monza), PSP (Pamplona), RIT (Rome), SIT (Cagliari), SSW (Stockholm), TIT (Trieste). i.e. centers with 'good' follow-up data and surgery data. The are the 'main centers'.

- Other centers (BIT, PCR, LPO, CAI, CEG, CRI, DEP, UDI, BAI, BCH, FIT, KPO, LIP, MCA, MFR, PFR, RZT, TUS, VAS) will be included in a supplementary analysis of performance in masses that are operated without follow-up (within 120 days after inclusion), together with the main centers. See below.

Exclusion criteria:

- Patients who did not give consent; data from patients who withdrew consent during the study will not be used either
- Younger than 18 years at first scan
- Patients who were already in follow-up at a center before the recruitment scan ('new' patients) (i.e. old patients)
- Cysts that are deemed clearly physiological if less than 3 cm in maximum diameter are not eligible for inclusion.

Multiple masses:

When the examiner detected multiple masses, the mass with the most complex ultrasound morphology was defined as the dominant mass. If multiple masses had similar morphology, the largest mass or the one best accessible with ultrasound was denoted as dominant.

Outcome:

Some patients received surgery without follow-up, some received follow-up only, some received follow-up with surgery later on, and for some patients we did not know what happened after the inclusion visit. We determine the dichotomous outcome (benign or malignant tumor) using the following rules:

A. If surgery, irrespective whether there was FU or not:
    - If the resulting histology was benign: classify as *benign (B1)*
    - If the resulting histology was malignant:
        - If surgery was within 120 days from the inclusion visit: classify as *malignant at inclusion (M1)*
        - If surgery was >120 days after the inclusion visit, but subjective assessment was probably or certainly malignant (including borderline) at every visit: classify as *malignant at inclusion (M2)*
        - If surgery was >120 days after the inclusion visit, but subjective assessment was not always probably or certainly malignant (including borderline) at every visit: classify as *uncertain (U1)*

B. The tumor spontaneously resolved at any point during follow-up: classify as *benign*

C. The patient did not receive surgery, the tumor did not resolve, and the last follow-up visit was >=10 months after the inclusion visit:

- If subjective assessment was consistently labeled as probably or certainly benign during the first 14 months of follow-up (specific definition: subjective assessment was benign at the earliest visit after >=10 months of follow-up, and subjective assessment was benign at all visits up to 14 months of follow-up): classify as *benign (B2)*
- If subjective assessment was consistently labeled as probably or certainly malignant (including borderline) during first 14 months of follow-up (specific definition: subjective assessment was borderline/malignant at the earliest visit after >=10 months of follow-up, and subjective assessment was borderline/malignant at all visits up to 14 months of follow-up): classify as *malignant (M3)*
- Else, subjective assessment was inconsistent/uncertain during the first 14 months of follow-up: classify as *uncertain (U2)*

D. Other cases: classify as *uncertain*
   - Patients that received follow-up of less than 10 months, without receiving surgery or experiencing spontaneous resolution of the tumor. Reasons for such short follow-up can be death, withdrawal from the study, or being lost to follow-up. (U3)
   - Patients for which we did not have information after the inclusion visit. Reasons for such short follow-up can be death, withdrawal from the study, or being lost to follow-up. (U4)

1.4 Models to be investigated

The following models will be validated:

- Risk of Malignancy Index (RMI) (Jacobs, 1990)
- IOTA LR2 model (Timmerman, 2005)
- IOTA Simple Rules (Timmerman 2008)
- IOTA Simple Rules Risk model (SRRisks) (Timmerman, 2016)
- IOTA ADNEX model (Van Calster, 2014)
- IOTA ADNEX without CA125 (Van Calster, 2014)

RMI, LR2, and SRRisks predict whether a tumor is malignant (vs benign), with borderline tumors being classified as malignant. ADNEX predicts malignant subgroups. More specifically, it gives 5 probabilities for each patient, for the following outcome categories: benign, borderline, stage I primary invasive, stage II-IV primary invasive, and secondary metastasis. One minus the probability of a benign tumor then equals the probability of malignancy.

In addition, we will also assess the performance of subjective assessment, which is customary in this field. We will assess subjective assessment as a binary judgment (benign vs malignant).

1.5 Statistical methods

1.5.1 Missing values for CA125 and outcome

CA125 was not mandatory in IOTA5, although it was highly encouraged. We expect missing values for CA125, for the following two reasons. First, following local protocols some centres were more committed than others to measure CA125. Second, the need to measure CA125 may depend on the overall clinical picture and the appearance of the tumour on ultrasound. Hence we expected clearly more missing values for cases that were followed up conservatively.

For some cases, as outlined above, the outcome is uncertain. The largest groups are cases without any information since the inclusion scan (U4 in the overview above), and cases with last FU <10 months after inclusion and without having had surgery and where the mass did not resolve spontaneously (U3).

Simply omitting cases with missing CA125 or outcome is likely to cause bias. For example, based on data from earlier IOTA studies, patients with missing CA125 have more often a benign tumor. The current data from IOTA5 indicate that CA125 is more often missing for patients that are followed conservatively than for patients that received surgery without follow-up visit. Further, we assume that conservatively followed patients with last follow-up <10 months will very often have a benign tumor.

Multiple imputation will be performed to deal with missing values for CA125 and outcome. Imputations will be created using the method of fully conditional specification with the mice package in R. We will generate 100 imputations, leading to 100 completed datasets. To estimate the missing values for CA125, we will use predictive mean matching regression using the outcome, variables that are probably related to either the level of CA125 itself, or to the unavailability of CA125 (i.e. a binary indicator indicating for each patient whether CA125 was missing). As the distribution of serum CA125 was heavily skewed, the log–log transformation of CA125 will be used (i.e., log(log(CA125 + 1))). In the imputation model, the following variables will be used: Presumed endometrioma, level of certainty at subjective assessment at inclusion (6 groups: certainly benign, probably benign, benign but uncertain, malignant but uncertain, probably malignant, certainly malignant), patient age, type of center (oncological versus non-oncological center), lesion largest diameter, proportion of solid tissue (maximum diameter solid component divided by maximum diameter of lesion), number of locules (1, 2-10, >10, other), number of papillations, presence of shadows, presence of ascites, presence of metastases, bilaterality, pelvic pain, personal history of ovarian cancer, papillary height, papillary flow, color score, echogenicity of cyst fluid, and outcome (benign, borderline, stage I primary invasive, stage II-IV primary invasive, secondary metastatic). Descriptive characteristics of the tumor were based on the inclusion scan. Uncertain outcomes (cf above) will be considered missing, and hence the outcome will be imputed).

Note that some patients are classified as having a malignant tumor based on clinical and ultrasound information during follow-up (cf category M3 above). For these patients, we do not have a classification into one of the malignancy subtypes. The multiple imputation procedure will therefore treat the outcome as missing such that it will be imputed. The imputed outcomes will be evaluated, and the most commonly imputed malignant type will be used as the outcome in the analysis.

1.5.2 How to deal with patients with uncertain outcome in the analysis

The outcome is a part of the imputation procedure described above, hence we have imputed values. The primary analysis is based on the multiply imputed values for CA125 and the outcome (cf de Groot, 2011).

As described below, we will also perform a sensitivity analysis after omitting patients with an uncertain outcome, and a sensitivity analysis after imputation of the outcome for patients where the outcome was labelled as uncertain, or was derived using information on subjective assessment (groups B2, M2, M3, U1, U2, U3, U4).

Note that models including CA125 (RMI and ADNEX with CA125) will always be analysed using imputed data for CA125, even when patients with uncertain outcome are omitted. Hence, only when evaluating models without CA125 based on patients for which the outcome is not labelled as uncertain, we do not have to use multiply imputed data.

1.5.3 Discrimination between benign and malignant tumors

We will calculate the c-statistic (i.e. AUC) per center, and calculate an overall AUC using random effects meta-analysis on the logit of the center-specific AUCs. In case multiple imputed data are used (e.g. models that include CA125, or when multiply imputed outcome are used), the logit(AUC) is calculated for each center and herein for each imputed dataset. The latter estimates will then be combined using Rubin's rules and the resulting estimates will be used in a random-effects model to get a final estimate of the AUC.

Sensitivity and specificity will be calculated at the following cut-offs (for models that give predicted risks): 1%, 3%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%. For the non-imputed data, a bivariate random-effects model will be used to calculate an overall result that takes the clustering by center into account. For imputed data, center-specific sensitivity and specificity as well as their variance (i.e. standard error squared) will be computed, and combined using Rubin's Rules to obtain a final center-specific estimate. These will be used in a bivariate random-effects model to come to a final estimate.

For RMI, we will use cut-offs of 25, 100, 200, and 250. For the Simple Rules classification, we will combine the inconclusive cases with the cases classified as malignant, i.e. inconclusive cases will be classified as malignant.

Because risk models and RMI are on a different scale, it is difficult to directly compare sensitivity or specificity. Therefore, we also calculate the sensitivity when fixing the specificity at 90%, and the specificity when fixing the sensitivity at 90%.

For subjective assessment, we will quantify the overall sensitivity and specificity using a bivariate random-effects model.

For Simple Rules, we will report the percentage of inconclusive cases, and specificity and sensitivity when inconclusives are classified as malignant. This will be done using meta-analysis techniques.

1.5.4. Calibration of the risk of malignancy

For non-imputed data, a random-effects logistic model will be used to compute both the calibration intercept and slope. The overall calibration curve will be computed using the results of the latter for an average center. The average center corresponds to a center where the calibration intercept and slope equal the fixed effect estimates (i.e. the random components are 0). Center-specific calibration curves will be computed using the empirical Bayes-estimates. For imputed data, the calibration intercept and slope will be computed per center and herein per imputed dataset by use of a logistic regression model.

Hereafter, the estimates will be combined using Rubin's Rules to get center-specific estimates and these estimates will be used in a random-effects model to get an overall estimate. For the overall calibration curve as well as for the center-specific calibration curves, this will be computed for each imputed dataset using the same method as for the non-imputed data and the results of the imputed datasets will be combined using Rubin's Rules.

The RMI does not give predicted risks. Therefore, similar to the IOTA3 validation study (Testa 2014), we will construct calibration curves conditional on the RMI score.

1.5.5. Clinical utility

We will calculate Net Benefit (NB) for risk thresholds between 5% and 50%, for using models to decide which patients to refer for specialized oncological care (Wynants 2017). For each center and threshold, we will make an average 2x2 cross-tabulation over the 100 imputed datasets. The cross-tabulation contrasts outcome (benign vs malignant) vs classification (risk<threshold vs risk≥threshold). This is used to calculate NB. Using meta-analysis, the center-specific NBs at a given threshold are combined into an overall estimate (Wynants, 2018).

1.5.6. Additional analysis for ADNEX

We will further analyse the ADNEX predictions in terms of discrimination and calibration. First, c-statistics between each pair of outcome categories will be computed using the conditional method (Van Calster, 2012). Only pooled results will be shown (i.e. no meta-analysis) because some outcome categories will have few events in several centers. For imputed data, logit(AUC) will be computed for each imputation, and combined using Rubin Rules.

Second, multinomial logistic calibration curves will be constructed (Van Hoorde 2014). For imputed data, curves will be derived per imputed dataset, and averaged.

1.5.7. Subgroup analyses

We will compute the overall c-statistic for benign vs malignancy (i.e. after meta-analysis) and overall calibration curves for the following prespecified subgroups:

- By actual management: separate analysis of patients who were operated within 120 days after the first scan and without follow-up scan (similar to the population for the IOTA1-4 studies), and patients who received at least 1 follow-up scan; note that not all recruited patients fall within one of these two categories
- By suggested management (‘intention to treat’): separate analysis of patients with suggested management surgery and patients with suggested management conservative
- By menopausal status: separate analysis of premenopausal patients and postmenopausal patient separately
- By type of center: separate analysis of patients examined in oncology centers and patients examined in other centers

The analysis of these subgroups will be done on data from the main centers, using meta-analysis if numbers allow. If numbers (e.g. number of patients with a malignancy) per center are too small, pooled analysis will be done.

No subgroup analyses are planned for the additional assessment of multinomial discrimination and calibration of the ADNEX model.

1.5.8. Sensitivity and supplementary analyses

The following sensitivity or supplementary analyses will be performed. For each analysis, we will report overall c-statistics for benign vs malignancy (i.e. after meta-analysis), as well as overall calibration curves.

- The 18 excluded centers (BIT, PCR, LPO, CAI, CEG, CRI, UDI, BAI, BCH, FIT, KPO, LIP, MCA, MFR, PCR, PFR, RZT, VAS) will be included in a supplementary analysis of performance in masses that are operated without follow-up (within 120 days after inclusion). This analysis will therefore include all centers: the 17 main centers listed above, and the 18 excluded centers. For this analysis, a separate multiple imputation procedure will be used.
- Sensitivity analysis where patients with an uncertain outcome as listed above (groups U1-U4 from section 1.3) are excluded rather than imputed.
- Sensitivity analysis after imputation of the outcome for all patients for which the outcome was uncertain or determined using subjective assessment (groups U1-U4, B2, N2, M3 from section 1.3).

1.6 References

De Groot JA, et al. Correcting for partial verification bias: a comparison of methods. Ann Epidemiol 2011;21:139-48.

Jacobs I, et al. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. BJOG 1990;97:922-9.

Testa A, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. Br J Cancer 2014;111:680-8.

Timmerman D, et al. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol 2000;16:500-5.

Timmerman D, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. J Clin Oncol 2005;23:8794-801.

Timmerman D, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol 2008;31:681-90.

Timmerman D, Van Calster B, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group. Am J Obstet Gynecol 2016;214:424-37.

Van Calster B, et al. 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 2014;349:g5920.

Van Calster B, et al. Assessing the discriminative ability of risk models for more than two outcome categories. Eur J Epidemiol 2012;27:761-70.

Van Hoorde K, et al. Assessing calibration of multinomial risk prediction models. Stat Med 2014;33:2585-96.

Wynants L, et al. Clinical Utility of Risk Models to Refer Patients with Adnexal Masses to Specialized Oncology Care: Multicenter External Validation Using Decision Curve Analysis. Clin Cancer Res 2017;23:5082-90.

Wynants L, et al. Random-effects meta-analysis of the clinical utility of tests and prediction models. Stat Med 2018;37:2034-52.